

(19)日本国特許庁 (J P)

(12) 公表特許公報 (A)

(11)特許出願公表番号

特表2001-511564

(P2001-511564A)

(43)公表日 平成13年8月14日(2001.8.14)

(51)Int.Cl.⁷

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/403

15/40

テーマコード^{*}(参考)

3 3 0 C 5 B 0 7 5

3 1 0 F

審査請求 未請求 予備審査請求 有 (全 94 頁)

(21)出願番号 特願2000-504525(P2000-504525)
(86) (22)出願日 平成10年5月13日(1998.5.13)
(85)翻訳文提出日 平成12年1月24日(2000.1.24)
(86)国際出願番号 PCT/US98/09711
(87)国際公開番号 WO99/05618
(87)国際公開日 平成11年2月4日(1999.2.4)
(31)優先権主張番号 08/898,652
(32)優先日 平成9年7月22日(1997.7.22)
(33)優先権主張国 米国 (US)
(81)指定国 EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), CN, JP

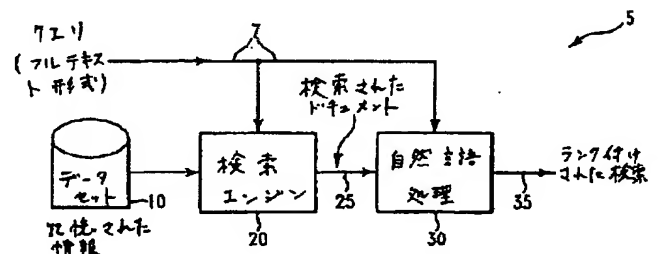
(71)出願人 マイクロソフト コーポレイション
Microsoft Corporation
アメリカ合衆国 ワシントン州 98052
レドモンド ワン マイクロソフト ウェイ (番地なし)
(72)発明者 ブレイデン-ハーダー, リサ
アメリカ合衆国、20194 パージニア州、
レストン、クリークベンド・ドライブ、
12003
(74)代理人 弁理士 深見 久郎 (外5名)

最終頁に続く

(54)【発明の名称】 全体の精度を高めるためにサーチ結果の自然言語処理を用いる情報検索システムのための装置および方法

(57)【要約】

全体の精度を高めるために、たとえば従来の統計に基づくサーチエンジンのような情報検索エンジンによって検索された結果を処理するために自然言語処理を利用する情報検索システムのための装置およびそれに付随する方法を提供する。具体的には、このようなサーチは最終的に検索されたドキュメントの集合を生む。このような各ドキュメントは次に自然言語処理を受けて論理形式の集合を生じる。このような各論理形式は句内の語間の意味的關係、特に主題と修飾語句との構造を語-関係子-語の態様で符号化する。ユーザが与えるクエリも同様に分析されてそのための対応の論理形式の集合を生み出す。ドキュメントはドキュメントおよびクエリからの論理形式の予め規定された関数としてランク付けされる。具体的には、クエリのための論理形式の集合は、検索されたドキュメントの各々のための論理形式の集合と比較されて両方の集合内のこのような任意の論理形式間の一致を確認する。少なくとも1つの一致する論理形式を有する各ドキュメントがヒューリスティックにスコア付けられ、一致する論理形式のための異なる各関係が異なる対



(2)

【特許請求の範囲】

【請求項 1】 記憶されているドキュメントをリポジトリから検索するための情報検索システムにおいて用いるための装置であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントを検索し、出力ドキュメント集合を規定するための検索システムを有し、前記装置は、プロセッサと、

実行可能な命令が記憶されているメモリとを含み、

プロセッサはメモリに記憶されている命令に応答して、

クエリに応答してそのための第 1 の論理形式を生じ、第 1 の論理形式はクエリに関連した語の間の意味的關係を示し、

出力ドキュメント集合内のドキュメントの各別の 1 つに対して、対応する第 2 の論理形式を取得し、第 2 の論理形式は前記 1 つのドキュメント内の句に関連した語の間の意味的關係を示し、

クエリの第 1 の論理形式と、出力ドキュメント集合内の複数のドキュメントの各 1 つのための第 2 の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定し、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与える、装置。

【請求項 2】 各エントリは出力ドキュメント集合内のドキュメントの対応の 1 つであるか、または前記対応の 1 つのドキュメントに関連したレコードである、請求項 1 に記載の装置。

【請求項 3】 クエリのための第 1 の論理形式と出力ドキュメント集合内の各別のドキュメントのための第 2 の論理形式との各々はそれぞれ、論理形式グラフ、そのサブグラフ、または論理形式三つ組のリストである、請求項 2 に記載の装置。

【請求項 4】 プロセッサは記憶されている命令に応答して、

出力ドキュメント集合内のドキュメントの前記各別の 1 つのために、記憶媒体から対応の第 2 の論理形式を読出すか、または

出力ドキュメント集合内の前記各別の 1 つのドキュメントを分析することによ

(3)

って前記対応の第 2 の論理形式を生成する、請求項 3 に記載の装置。

【請求項 5】 前記関数は、前記ドキュメントの 1 つのために、クエリに関連した前記第 1 の論理形式と前記 1 つのドキュメントに関連した前記第 2 の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、プロセッサは記憶されている命令に応答して、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定する、請求項 4 に記載の装置。

【請求項 6】 クエリに関連した前記第 1 の論理形式または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した前記第 2 の論理形式は、それぞれ前記クエリにまたは前記ドキュメントの 1 つに関連した語句のパラフレーズをさらに含む、請求項 5 に記載の装置。

【請求項 7】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の、対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、 2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 6 に記載の装置。

【請求項 8】 クエリに関連した前記第 1 の論理形式と出力ドキュメント集合内の任意のドキュメントに関連した前記第 2 の論理形式の任意のものとの間の前記一致は同一の一致である、請求項 5 に記載の装置。

【請求項 9】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、 2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 8 に記載の装置。

【請求項 10】 リポジトリはデータセットを含む、請求項 5 に記載の装置

(4)

。

【請求項 1 1】 クエリはフルテキストのクエリである、請求項 5 に記載の装置。

【請求項 1 2】 検索システムは統計的サーチエンジンを含む、請求項 5 に記載の装置。

【請求項 1 3】 ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライアントコンピュータと、

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライアントコンピュータからクエリを取得し、

出力ドキュメントの集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与える、請求項 5 に記載の装置。

【請求項 1 4】 前記サーバは複数の個別のサーバを含む、請求項 1 3 に記載の装置。

【請求項 1 5】 検索システムは統計的サーチエンジンを含む、請求項 1 3 に記載の装置。

【請求項 1 6】 ネットワーク接続はインターネットまたはインターネット接続である、請求項 1 5 に記載の装置。

【請求項 1 7】 サーチエンジンはクエリに応答して、出力ドキュメントの集合内の前記複数のドキュメントの各 1 つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項 1 6 に記載の装置。

【請求項 1 8】 前記プロセッサおよび前記メモリを有するクライアントコンピュータと、

(5)

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与える、請求項5に記載の装置。

【請求項19】 検索システムは統計的サーチエンジンを含む、請求項18に記載の装置。

【請求項20】 ネットワーク接続はインターネットまたはインターネット接続である、請求項19に記載の装置。

【請求項21】 サーチエンジンはクエリに応答して、出力ドキュメントの集合内の前記複数のドキュメントの各1つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項20に記載の装置。

【請求項22】 前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応じて前記検索システムを実現する、請求項5に記載の装置。

【請求項23】 検索システムは統計的サーチエンジンを含む、請求項22に記載の装置。

【請求項24】 前記1つのドキュメントのためのスコアはまた、前記1つのドキュメントのための第2の論理形式内のノード語、前記1つのドキュメント内の前記ノード語の頻度または意味的内容、前記1つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記1つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記1つのドキュメントの長さの、予め定められた関数である、請求項5に記載の装置。

【請求項25】 クエリはフルテキストのクエリである、請求項24に記載の装置。

【請求項26】 検索システムは統計的サーチエンジンを含む、請求項24

(6)

に記載の装置。

【請求項27】 ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライアントコンピュータと、

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライアントコンピュータからクエリを取得し、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与える、請求項24に記載の装置。

【請求項28】 サーバは複数の別個のサーバを含む、請求項27に記載の装置。

【請求項29】 検索システムは統計的サーチエンジンを含む、請求項27に記載の装置。

【請求項30】 ネットワーク接続はインターネットまたはインターネット接続である、請求項29に記載の装置。

【請求項31】 サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応じて、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項30に記載の装置。

【請求項32】 前記プロセッサおよび前記メモリを有するクライアントコンピュータと、

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与える、請求項24に記載の装置。

(7)

【請求項 3 3】 検索システムは統計的サーチエンジンを含む、請求項 3 2 に記載の装置。

【請求項 3 4】 ネットワーク接続はインターネットまたはインターネット接続である、請求項 3 3 に記載の装置。

【請求項 3 5】 サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各 1 つのためにリポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項 3 4 に記載の装置。

【請求項 3 6】 前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応答して前記検索システムを実現する、請求項 2 4 に記載の装置。

【請求項 3 7】 検索システムは統計的サーチエンジンを含む、請求項 3 6 に記載の装置。

【請求項 3 8】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、 2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 5 に記載の装置。

【請求項 3 9】 クエリに関連した論理形式三つ組の前記第 1 のリストか、または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した論理形式三つ組の前記第 2 のリストは、それぞれ前記クエリにまたは前記ドキュメントの 1 つに関連した語句のパラフレーズをさらに含む、請求項 3 8 に記載の装置。

【請求項 4 0】 前記 1 つのドキュメントのためのスコアはまた、前記 1 つのドキュメントのための第 2 の論理形式内のノード語、前記 1 つのドキュメント

(8)

内の前記ノード語の頻度または意味的内容、前記 1 つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記 1 つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記 1 つのドキュメントの長さの、予め定められた関数である、請求項 38 に記載の装置。

【請求項 41】 前記関数は、クエリに関連した論理形式三つ組の少なくとも 1 つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的關係のタイプによって定義される、請求項 38 に記載の装置。

【請求項 42】 プロセッサはメモリに記憶されている命令に応答して、クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定し、

関連した少なくとも 1 つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各 1 つのために、前記一致する論理形式三つ組の各々に関連した意味的關係によって予め規定される重み数値を用いて前記各 1 つのドキュメント内の一致する論理形式三つ組に重み付けして、前記 1 つのドキュメントのための 1 つ以上の重みを形成し、

前記 1 つ以上の重みの関数として前記 1 つのドキュメントのためのスコアを計算し、

前記ドキュメントの各 1 つをその前記スコアに従ってランク付けしてランク順を規定する、請求項 41 に記載の装置。

【請求項 43】 ランク順は重みの大きいものから小さいものの順である、請求項 42 に記載の装置。

【請求項 44】 プロセッサはメモリに記憶されている命令に応答して、前記出力ドキュメント集合内のドキュメントの、最も高い、連続するランク付けを有する、前記出力ドキュメント集合のための前記エントリの第 1 の予め規定されたグループを提示する、請求項 38 に記載の装置。

【請求項 45】 出力ドキュメント集合内の複数のドキュメントは、関連し

(9)

た少なくとも 1 つの一致する三つ組を有する、前記出力ドキュメント集合内のドキュメントからなる、請求項 4 4 に記載の装置。

【請求項 4 6】 前記第 1 の論理形式三つ組および前記第 2 の論理形式三つ組の各々は、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 4 5 に記載の装置。

【請求項 4 7】 クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項 3 8 に記載の装置。

【請求項 4 8】 クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項 3 8 に記載の装置。

【請求項 4 9】 リポジトリはデータセットを含む、請求項 3 8 に記載の装置。

【請求項 5 0】 クエリはフルテキストのクエリである、請求項 3 8 に記載の装置。

【請求項 5 1】 検索システムは統計的サーチエンジンを含む、請求項 3 8 に記載の装置。

【請求項 5 2】 ユーザからのクエリを取得し、出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するためのクライアントコンピュータと、

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記プロセッサおよび前記メモリを含み、プロセッサはメモリに記憶されている命令に応答して、

クライアントコンピュータからクエリを取得し、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与える、請求項 3 8 に記載の装置。

(10)

【請求項53】 サーバは複数の個別のサーバを含む、請求項52に記載の装置。

【請求項54】 検索システムは統計的サーチエンジンを含む、請求項52に記載の装置。

【請求項55】 ネットワーク接続はインターネットまたはインターネット接続である、請求項54に記載の装置。

【請求項56】 サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのために、リポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項55に記載の装置。

【請求項57】 前記プロセッサおよび前記メモリを有するクライアントコンピュータと、

ネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与える、請求項38に記載の装置。

【請求項58】 検索システムは統計的サーチエンジンを含む、請求項57に記載の装置。

【請求項59】 ネットワーク接続はインターネットまたはインターネット接続である、請求項58に記載の装置。

【請求項60】 サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのために、リポジトリから記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力

(11)

ドキュメント集合内に含める、請求項 5 9 に記載の装置。

【請求項 6 1】 前記プロセッサおよび前記メモリを有するコンピュータをさらに含み、コンピュータはまたメモリに記憶されている命令に応答して前記検索システムを実現する、請求項 3 8 に記載の装置。

【請求項 6 2】 検索システムは統計的サーチエンジンを含む、請求項 6 1 に記載の装置。

【請求項 6 3】 記憶されているドキュメントをリポジトリから検索するための情報検索システムにおいて用いるための方法であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントを検索し、出力ドキュメント集合を規定するための検索システムを有し、前記方法は、クエリに応答してそのための第 1 の論理形式を生じるステップを含み、第 1 の論理形式はクエリに関連した語の間の意味的關係を示し、

出力ドキュメント集合内のドキュメントの各別の 1 つに対して、対応する第 2 の論理形式を取得するステップを含み、第 2 の論理形式は前記 1 つのドキュメント内の句に関連した語の間の意味的關係を示し、

クエリの第 1 の論理形式と、出力ドキュメント集合内の複数のドキュメントの各 1 つのための第 2 の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定するステップと、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与えるステップとを含む、方法。

【請求項 6 4】 各エントリは出力ドキュメント集合内のドキュメントの対応の 1 つであるか、または前記対応の 1 つのドキュメントに関連したレコードである、請求項 6 3 に記載の方法。

【請求項 6 5】 クエリのための第 1 の論理形式と出力ドキュメント集合内の各別のドキュメントのための第 2 の論理形式との各々はそれぞれ、論理形式グラフ、そのサブグラフ、または論理形式三つ組のリストである、請求項 6 4 に記載の方法。

【請求項 6 6】 前記取得するステップは、

(12)

出力ドキュメント集合内のドキュメントの前記各別の 1 つのために、記憶媒体から対応の第 2 の論理形式を読出すか、または

出力ドキュメント集合内の前記各別の 1 つのドキュメントを分析することによって、前記対応の第 2 の論理形式を生成するステップを含む、請求項 6 5 に記載の方法。

【請求項 6 7】 前記関数は、前記ドキュメントの 1 つのために、クエリに関連した前記第 1 の論理形式と前記 1 つのドキュメントに関連した前記第 2 の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、前記ランク付けするステップは、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定するステップを含む、請求項 6 6 に記載の方法。

【請求項 6 8】 クエリに関連した前記第 1 の論理形式または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した前記第 2 の論理形式は、それぞれ前記クエリにまたは前記ドキュメントの 1 つに関連した語句のパラフレーズをさらに含む、請求項 6 7 に記載の方法。

【請求項 6 9】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の、対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 6 8 に記載の方法。

【請求項 7 0】 クエリに関連した前記第 1 の論理形式の任意のものと出力ドキュメント集合内の任意のドキュメントに関連した前記第 2 の論理形式の任意のものとの間の前記一致は同一の一致である、請求項 6 7 に記載の方法。

【請求項 7 1】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の、対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、

(13)

対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、2 つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項 7 0 に記載の方法。

【請求項 7 2】 リポジトリはデータセットを含む、請求項 6 7 に記載の方法。

【請求項 7 3】 クエリはフルテキストのクエリである、請求項 6 7 に記載の方法。

【請求項 7 4】 検索システムは統計的サーチエンジンを含む、請求項 6 7 に記載の方法。

【請求項 7 5】 システムはクライアントコンピュータをさらに含み、前記方法はクライアントコンピュータにおいて、

ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとを含み、

システムはネットワーク接続を介してクライアントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライアントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与えるステップとを含む、請求項 6 7 に記載の方法。

【請求項 7 6】 検索システムは統計的サーチエンジンである、請求項 7 5 に記載の方法。

【請求項 7 7】 ネットワーク接続はインターネットまたはインターネット接続である、請求項 7 6 に記載の方法。

【請求項 7 8】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各 1 つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出される得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、

(14)

出力ドキュメント集合内に含めるステップをさらに含む、請求項77に記載の方法。

【請求項79】 システムはクライアントコンピュータとネットワーク接続を介してクライアントコンピュータに接続されるサーバとをさらに含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与えるステップをさらに含む、請求項67に記載の方法。

【請求項80】 検索システムは統計的サーチエンジンを含む、請求項79に記載の方法。

【請求項81】 ネットワーク接続はインターネットまたはインターネット接続である、請求項80に記載の方法。

【請求項82】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメントの集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、クライアントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項81に記載の方法。

【請求項83】 システムはコンピュータをさらに含み、前記方法はコンピュータにおいて前記検索システムを実現するステップを含む、請求項67に記載の方法。

【請求項84】 検索システムは統計的サーチエンジンを含む、請求項83に記載の方法。

【請求項85】 前記1つのドキュメントのためのスコアはまた、前記1つのドキュメントのための第2の論理形式内のノード語、前記1つのドキュメント内の前記ノード語の頻度または意味的内容、前記1つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記1つのドキュメントのための特

(15)

定の論理形式三つ組の頻度、もしくは前記 1 つのドキュメントの長さの、予め定められた関数である、請求項 6 7 に記載の方法。

【請求項 8 6】 リポジトリはデータセットを含む、請求項 8 5 に記載の方法。

【請求項 8 7】 クエリはフルテキストのクエリである、請求項 8 5 に記載の方法。

【請求項 8 8】 検索システムは統計的サーチエンジンを含む、請求項 8 5 に記載の方法。

【請求項 8 9】 システムはクライアントコンピュータをさらに含み、前記方法はクライアントコンピュータにおいて、

ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとをさらに含み、

システムはネットワーク接続を介してクライアントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライアントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与えるステップとをさらに含む、請求項 8 5 に記載の方法。

【請求項 9 0】 検索システムは統計的サーチエンジンを含む、請求項 8 9 に記載の方法。

【請求項 9 1】 ネットワーク接続はインターネットまたはインターネット接続である、請求項 9 0 に記載の方法。

【請求項 9 2】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各 1 つのためにリポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項 9 1 に記載の方法

(16)

。

【請求項 9 3】 システムはクライアントコンピュータとネットワーク接続を介してクライアントコンピュータに接続されるサーバとを含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与えるステップをさらに含む、請求項 8 5 に記載の方法。

【請求項 9 4】 検索システムは統計的サーチエンジンを含む、請求項 9 3 に記載の方法。

【請求項 9 5】 ネットワーク接続はインターネットまたはインターネット接続である、請求項 9 4 に記載の方法。

【請求項 9 6】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各 1 つのためにリポジトリから記憶されているレコードを検索するステップをさらに含む、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出され得る場所を特定する情報を含み、クライアントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項 9 5 に記載の方法。

【請求項 9 7】 システムはコンピュータをさらに含む、前記方法はコンピュータにおいて前記検索システムを実現するステップを含む、請求項 8 5 に記載の方法。

【請求項 9 8】 検索システムは統計的サーチエンジンを含む、請求項 9 7 に記載の方法。

【請求項 9 9】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、 2 つの語の間の意味的関係を表わす予め規定された関係とを含む、請求項 6 7 に記

(17)

載の方法。

【請求項 1 0 0】 クエリに関連した論理形式三つ組の前記第 1 のリストか、または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した論理形式三つ組の前記第 2 のリストは、それぞれ前記クエリにまたは前記ドキュメントの 1 つに関連した語句のパラフレーズをさらに含む、請求項 9 9 に記載の方法。

【請求項 1 0 1】 前記 1 つのドキュメントのためのスコアはまた、前記 1 つのドキュメントのための第 2 の論理形式内のノード語、前記 1 つのドキュメント内の前記ノード語の頻度または意味的内容、前記 1 つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記 1 つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記 1 つのドキュメントの長さの、予め定められた関数である、請求項 9 9 に記載の方法。

【請求項 1 0 2】 前記関数は、クエリに関連した論理形式三つ組の少なくとも 1 つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的關係のタイプによって定義される、請求項 9 9 に記載の方法。

【請求項 1 0 3】 前記ランク付けするステップは、クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定するステップと、

関連した少なくとも 1 つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各 1 つのために、前記一致する論理形式三つ組の各々に関連した意味的關係によって予め規定される重み数値を用いて前記各 1 つのドキュメント内の一致する論理形式三つ組に重み付けして、前記 1 つのドキュメントのための 1 つ以上の重みを形成するステップと、

前記 1 つ以上の重みの関数として前記 1 つのドキュメントのためのスコアを計算するステップと、

前記ドキュメントの各 1 つをその前記スコアに従ってランク付けしてランク順

(18)

を規定するステップとを含む、請求項102に記載の方法。

【請求項104】 ランク順は重みの大きいものから小さいものの順である、請求項103に記載の方法。

【請求項105】 記憶されているエントリを与えるステップは、前記出力ドキュメント集合内のドキュメントの、最も高い、連続するランク付けを有する、前記出力ドキュメント集合のための前記エントリの第1の予め規定されたグループを提示するステップを含む、請求項99に記載の方法。

【請求項106】 出力ドキュメント集合内の前記複数のドキュメントは、関連した少なくとも1つの一致する三つ組を有する、前記出力ドキュメント集合内のドキュメントからなる、請求項105に記載の方法。

【請求項107】 前記第1の論理形式三つ組および前記第2の論理形式三つ組の各々は、それぞれクエリ内のまたは前記ドキュメントの各1つの句内の、対応の論理形式グラフにおいて意味的に関係した2つの語の各々の語幹の形と、2つの語の間の意味的关系を表わす予め規定された関係とを含む、請求項106に記載の方法。

【請求項108】 クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの1つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項99に記載の方法。

【請求項109】 クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項99に記載の方法。

【請求項110】 リポジトリはデータセットを含む、請求項99に記載の方法。

【請求項111】 クエリはフルテキストのクエリである、請求項99に記載の方法。

【請求項112】 検索システムは統計的サーチエンジンを含む、請求項99に記載の方法。

【請求項113】 前記システムはクライアントコンピュータをさらに含み

(19)

、前記方法はクライアントコンピュータにおいて、
ユーザからのクエリを取得するステップと、

出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示するステップとを含み、

前記システムはネットワーク接続を介してクライアントコンピュータに接続されるサーバをさらに含み、前記方法はサーバにおいて、

クライアントコンピュータからクエリを取得するステップと、

出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与えるステップとをさらに含む、請求項99に記載の方法。

【請求項114】 検索システムは統計的サーチエンジンを含む、請求項113に記載の方法。

【請求項115】 ネットワーク接続はインターネットまたはインターネット接続である、請求項114に記載の方法。

【請求項116】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのために、リポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各1つにそのための関連するサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項115に記載の方法。

【請求項117】 システムはクライアントコンピュータとネットワーク接続を介してクライアントコンピュータに接続されるサーバとを含み、前記サーバは前記検索システムを実現し、前記方法は、サーバにおいて、クライアントコンピュータによって与えられるクエリに応答して前記出力ドキュメント集合をクライアントコンピュータに与えるステップをさらに含む、請求項99に記載の方法。

【請求項118】 検索システムは統計的サーチエンジンを含む、請求項117に記載の方法。

【請求項119】 ネットワーク接続はインターネットまたはインターネット接続である、請求項118に記載の方法。

【請求項120】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのために、リポジトリから記憶されているレコードを検索するステップをさらに含み、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、クライアントコンピュータにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各1つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項119に記載の方法。

【請求項121】 システムはコンピュータをさらに含み、前記方法はコンピュータにおいて前記検索システムを実現するステップをさらに含む、請求項9に記載の方法。

【請求項122】 検索システムは統計的サーチエンジンを含む、請求項121に記載の方法。

【請求項123】 コンピュータで実行可能な命令を記憶し、請求項63に記載のステップを実行するためのコンピュータ読出可能媒体。

(21)

【発明の詳細な説明】**【0001】****【発明の分野】**

本発明は、たとえば統計に基づいた従来のサーチエンジンなどの情報検索エンジンによって検索された結果を処理するために自然言語処理を利用して全体的な精度を改善する、情報検索システムのための装置およびそれに付随する方法に関する。

【0002】**【先行技術の説明】**

数十年から現在に至るまで、自動情報検索技術は、出版物および／またはそのための書誌情報を含む従来のデータベースなどの大容量データ記憶装置から記憶情報を検索するためにますます頻繁に使用されている。このような従来のデータベースはたとえば米国電気電子通信学会（IEEE）によって維持されており、かつたとえばKnight-Ridder Information Inc.のダイアログ（Dialog）情報サービスによって現在アクセス可能である、INSPECデータベースの場合のように、電気工学およびコンピュータ関連技術などの、広い範囲ではあるが特定のトピックに向けられる情報を一般的に含むため、専門的になる傾向がある（DIALOGはKnight-Ridder Information Inc.の登録サービスマークである）。このタイプのデータベースは明らかに、関連の論文および他の出版物の出版数の増加に伴い増大し続けることは確実だが、この増大は比較的なだらかで、適度にうまく調整される傾向がある。さらに、このように専門化されたデータベースは比較的うまく組織化されやすい。

【0003】

しかしながら、インターネットを介してアクセス可能ないわゆる「ワールドワイドウェブ」（以下単に「ウェブ」と称する）の出現および隆盛により、また従来の出版とは対照的にウェブに情報を投稿しかつそれから情報にアクセスする、比較的容易で低コストで行えることにより、ウェブ上で利用できる情報量は、爆発的ではないにしても非常に指数関数的に近い増加を遂げ、現実的な視界内に制限があるようには思われない。ウェブはおよそ人間が試みる全ての学問分野に及

ぶ、増々膨大な量の情報を提供するが、ウェブ上の情報内容は非常に無秩序であり、極端な程に組織化されておらず、そのためウェブからの情報のアクセスおよび検索を非常に複雑にし、多大な労苦を強いる。

【0004】

ウェブからの情報の検索作業をはるかに容易にするために、過去数年にコンピュータによる多数のサーチエンジンが開発されており、広く一般的に使用されている。概して、これらの従来のエンジンは、ソフトウェアにより実装された「ウェブクローラ」によって、ウェブサイトに自動的に入り、その中のハイパーテキストリンクを順次追跡し、その中にある各ドキュメントを抽出し、要約し、かつそれにいわゆる「キーワード」によって索引を付けて大型データベースとし、後にアクセスできるようにする。具体的には、このような要約により、クローラが遭遇するこのような各ドキュメントは通常「1袋の単語」と呼ばれるものにまで凝縮される。これは、意味論上および統語論上の全ての情報は除去されているが、ドキュメント内に存在する内容語を含む。内容語は、ドキュメント自体にあることもあるし、および／またはそのドキュメントのハイパーテキストマークアップ言語（HTML）版の記述フィールドにだけ現われることもある。いずれにせよ、エンジンは、そうしたドキュメントに対するエントリ、すなわちドキュメントレコードを作成する。各ドキュメントについて、その内容語の各々に索引が付けられて、そのドキュメントに戻るリンクを有する、サーチ可能なデータ構造が形成される。ドキュメントレコードは典型的には、（a）ウェブアドレス、すなわちURL—これによってウェブブラウザによって対応のドキュメントをアクセスすることができるユニフォームリソースロケータと、（b）そのドキュメント内の、さまざまな内容語とを含み、さらに、エンジンによっては、そのドキュメント内の他の内容語に対する、これら内容語の各々の相対的なアドレスを含み、さらに（c）概要とを含み、これはドキュメントのうち数行のみであるか、またはドキュメントの最初の数行であることが多いが、さらに場合によってはd）そのHTMLの記述フィールド内に記載された、そのドキュメントに関する説明とを含む。データベースをサーチするために、ユーザはキーワードに基づいたクエリをエンジンに与える。

【0005】

クエリは典型的には、ユーザによって与えられた1または2以上、多くの場合には小さな数だけのキーワードを含み、エンジンの能力に依存して、場合によっては連続したキーワード間にあるブール論理（たとえば「AND」または「OR」）、または類似したオペレータ（たとえば数的近さ）を含む。クエリに応答して、エンジンはできるだけ多くのキーワードと、論理的なまたは近さに関するオペレータが提供されている場合には、リクエストされた特定の組合わせまたはお互いにある「レンジ」（特定の数の、内容語）内にあるようなキーワードとを含むドキュメントを突き止めようとする。これを行なう際に、エンジンはそのデータベースをサーチして、クエリのキーワードの1つと一致する少なくとも1つの単語を含み、リクエストがあつた場合にはそのリクエストによって特定されたオペレータおよび／またはレンジに一致するような、ドキュメントを突き止める。エンジンは見つけ出したこのようなドキュメントの各々に対して、それに関するドキュメントレコードを検索し、そのドキュメント内のキーワード一致の数に従って、同様の他のこうした文書に対するランク付けをしてそのレコードをユーザに提示する。

【0006】

ユーザによって与えられるキーワードのクエリに応答して検索されただけのドキュメントの大部分はクエリとは全く関連がないことが多く、ユーザをいらだたせる。

【0007】

したがって、無関係なドキュメントが検索される数を減らすために、キーワードに基づいた従来のサーチエンジン（以下、単に「統計的サーチエンジン」と称する）は、統計的处理をそれらのサーチ法に取り入れている。たとえば、クエリ内のキーワードと検索された各ドキュメントレコード中の内容語との間で一致するキーワードの総数およびこれらの単語の一致の程度、すなわち組合せとしておよび／またはリクエストされた近さのレンジ内にあるか否かに基づいて、統計的サーチエンジンは、検索されたこのようなドキュメントレコードの各々に対して、「統計値」と包括的に呼ばれることが多い数値尺度を計算する。これらの統計

(24)

値は、一致する各単語に対するドキュメント頻度の逆数を含み得る。その後エンジンはそれらの統計値によってドキュメントレコードをランク付けし、最もランクの高い、典型的には5個から20個以下の、予め規定された少数の検索済レコードに関するドキュメントレコードを、ユーザに戻す。検索された第1のグループのドキュメントに対するユーザが第1のグループのドキュメントレコード（またはある種のエンジンのように、エンジンによってドキュメントが戻される場合にはドキュメント自体）をユーザが検討すると、ユーザは、次に高いランキングのドキュメントレコードのグループを要求することができ、以下同様に検索されたドキュメントレコード全てがこうして検討されるまで、要求をすることができる。

【0008】

従来、サーチエンジンの性能は再現および精度によって評価されてきた。再現は、データセット内の関連の全てのドキュメントに対し、所与のクエリに応答して実際に検索されたこのようなドキュメントの数を百分率で測る。一方、精度は、検索された全てのドキュメントに対し、クエリに実際に関連するドキュメントの数を百分率で測る。最終的に検索されるドキュメントの単なる数は重要ではないため、ウェブサーチエンジンについて考える場合には、再現は性能に関する重要な測定基準ではないと考える。実際に、クエリによってはこの数は過度に大きいこともある。したがって、有用な結果を生み出すためには、エンジンによって索引付けされた関連のドキュメントの全てを取り出す必要はないと考える。しかしながら精度は極めて重要であると考えられる。すなわち、ランクが最も高く最初にユーザに提示されるドキュメントは、クエリに最も関連するものであるべきであると考ええる。

【0009】

従来の統計的サーチエンジンの精度が比較的低いのは、単語が独立変数である、すなわち全ての文章の単語は互いに独立して現われるという仮定に基づいていることに由来する。この場合の独立とは、別のある単語が文書中に存在するときにそのドキュメントに任意の1つの単語が表れるという条件付確率が常にゼロであること、すなわち、ドキュメントが、構造を持たない単語の集まりを含むだけ

であるか、または単に「1袋の単語」でしかないことを意味する。容易に認識できるように、この仮定は全ての言語に関して非常に誤ったものである。英語は、他の言語と同様に、単語に関する、膨大な量で複雑な統語論上および語彙一意味論上の構造を有し、これらの単語の意味は、使用される特定の言語的文脈に基づいて、しばしば広く、異なることが多く、文脈はその場合にも単語に与えられた意味と、いかなる単語が後に現われるかとを決定する。したがって、文章に表れる単語は単に独立しているのでは全くなく、相互に高度に依存する。キーワードに基づいたサーチエンジンはこのきめの細かな言語的構造を全く無視している。たとえば、自然言語で表現された「How many hearts does an octopus have?」というクエリの例について考える。内容語「hearts」および「octopus」またはその形態素的語幹に基づいて動作する統計的サーチエンジンであれば、その材料の部分に、したがってその内容を表わす単語として「artichoke hearts(アーチチョークの芯)からsquid(イカ)、onions(タマネギ)およびoctopus(タコ)」を有するレシピを含む、記憶されたドキュメントを戻すか、またはユーザをそのドキュメントに導く。内容を表わす2つの内容単語「octopus」および「hearts」が一致するので、このエンジンは、たとえば近さおよび論理的なオペレータを含む統計的測定値に基づいて、実際にはドキュメントがクエリとはかなり無関係であっても、このドキュメントが優れた一致であると決定してしまう。

【0010】

この技術分野では、ラベル付きでない関係にあるヘッダー修正語対として統語論的な句の要素を抽出するためのさまざまな方策が教示されている。これらの要素はその後、従来の統計的ベクトル空間モデルにおける（典型的には内部構造のない）述語として索引付けされる。

【0011】

このような方策の一例はJ. L. Fagan, "Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods", Ph.D. Thesis, Cornell University, 1988, pages i-261に記載されている。具体的には、この方策は自然言語処理を使用して英語の文章を分析して、統語論的な句の構成要素を抽出し、これらの句の構成要素は後に術語として

扱われ、統計的なベクトル空間モデルを用いた索引に索引付けされる。検索時には、ユーザは自然言語でクエリを入力し、この方策の下では、このクエリに自然言語処理が施され、分析され、索引中検索にレコードされている憶された要素に類似した統語論的な句の構成要素の要素を抽出する。その後、クエリによる統語論的な句の構成要素と、索引に記憶されているものとのを比較照合一致が行なわれるせる試みがなされる。著者は、この純粹に統語論的な方策を、統語論的な句の構成要素を特定するために確率論的方法を使用すると統計的な方策とを対比させており、この統計的な方策では統語論的な句の構成要素を特定するために確率論的方法が使用される。著者は、自然言語処理では確率論的な方策からの大幅な実質的に改善は見られず、自然言語処理が時に提供し得る精度がのわずかに軽微な改善されることによつては、自然言語処理に伴うかなりの処理コストはを正当化するものではない、と結論し付けている。

【0012】

サーチ用のクエリに含ませるための適切な単語を選択するための自然言語処理を使用する場合の、このような統語論に基づいた別の方策が、T. Strzalkowski, "Natural Language Information Retrieval: TIPSTER-2 Final Report", Proceedings of Advances in Text Processing: Tipster Program Phase 2, DARPA, 6-8 May 1996, Tysons Corner, Virginia, pages 143-148 (以下、「DARPAの論文」と呼ぶ) および T. Strzalkowski, "Natural Language Information Retrieval", Information Processing and Management, Vol. 31, No. 3, 1995, pages 397-417に記載されている。この方策は理論的には可能性を秘めたものではあるが、DARPAの論文の第147頁から148頁において著者は、基盤となる自然言語技術を実現するために要求される処理が高度であるために、この方策は現在のところは実用的でない、と結論している。

「...ただし、我々の性能要件を満たす（またはこれらの要件に少なくとも近いと考えられる）NLP〔自然言語処理〕技術は、依然として、自然言語の文を扱う能力において性能がかなり低い。特に、概念的構成および論理形式などにかかわる進歩した処理には、計算の面から依然として届かない。これらの進歩した技術は表現レベルの限界の問題に対処するものであるため、より有効となるであ

(27)

うと仮定することもできるが、実証に乏しく、しかもかなり小さなスケールのテストに限定されざるを得ない。」

この種の統合論に基づいたさらなる方策がB. Katz, “Annotating the World Wide Web Using Natural Language”, Conference Proceedings of RIAO 97, Computer-Assisted Information Searching in Internet, McGill University, Quebec, Canada, 25-27 June 1997, Vol. 1, pages 136-155 (以下、「Katzの出版物」と称する)に記載されている。Katzの出版物の記載によれば、内部構造を維持したままで主語—動詞—目的語表現が作成されるので、検索時には軽微な統語論上の変更に対処できる。

【0013】

これらの統語論的方策が大した改善をもたらさなかったこと、またはその時点で利用可能な自然言語処理システムによっては実現できなかったことにより、研究分野は、クエリによる最初の結果の精度および再現を直接改善しようとする試みから、ユーザインターフェイスの改善へ、すなわち具体的には、「類語を検索せよ」とユーザが検索結果に応答することなどの、ユーザとの相互作用に基づいてクエリの精度を向上させる方法、および適当な固まりに分けて結果を表示することを含む、クエリに対する結果を視覚化するための方法による改善に、移行した。

【0014】

これらの改善自体は有用であるが、これらの改善によって達成可能な精度の向上は依然として失望するほど少なく、キーワードによるサーチに特有なユーザの感ずる歯がゆさを大幅に軽減するには明らかに不十分である。具体的には、ユーザには、関連ある応答がまばらにしかないような比較的大きなドキュメントの集合を手操作によってふるいにかけることが依然として要求される。

【0015】

したがって、この技術分野においては、情報検索に対する従来の統計的方策によって達成可能なものに優る著しい精度の改善をもたらすことができる、情報を検索するための技術、特定的には装置およびそれに付随する方法が必要とされている。さらに、このような技術は、任意に生ずる文章内の、広範囲な文のタイプ

および長さに対しても信頼性が高くかつ反復可能な結果をもたらし、かつ実用的であり実現の際のコスト面でも有効である必要がある。このような従来の方策の精度に対して、しかもこの技術分野に特有な問題にもかかわらず、精度を著しく改善するためには、このような技術は好ましくは、意味的な内容とクエリの内容との照合に基づいて、関連のドキュメントを選択し検索してその後ユーザに提示するという効果を得るために自然言語処理を利用すべきである。

【0016】

【発明の概要】

我々の広い教示によると、本発明は、たとえば統計的ウェブサーチエンジンによって行なわれるキーワードに基づいたドキュメントサーチの精度を改善するために自然言語処理を採用することにより、この必要を満たす。

【0017】

大まかに言って、この処理は、それぞれがサーチクエリおよび検索されたドキュメントの各々に関連した論理形式を生成し、比較し、それらの一致を重み付けすることを含む。検索されたドキュメントは、クエリおよび検索されたドキュメントの両方に関する「論理形式」の予め規定された関数に基づいて、特定的には、ドキュメントに関連した一致する論理形式に関連する重みの和に基づいて、ランク付けされ、その順で最終的に表示される。論理形式とは、任意のサイズのテキストを表わす単語がラベル付きの関係によってリンクされる有向非循環グラフである。特に、論理形式は入力文字列における重要な単語間の意味論上の関係を、特に主題および修飾語句の関係を描く。この描写はさまざまな特定の形式をとることができ、たとえば論理形式グラフまたは、たとえば論理形式三つ組 (triple) のリストを含むその任意のサブグラフの形式をとることができる。ここで、三つ組の各々はたとえば「語－関係詞－語」という形式をとるが、これら形式のいずれでも我々の発明に用いることができる。

【0018】

我々の特定の教示によると、このようなサーチにより最終的には、たとえばデータベースまたはワールドワイドウェブからの検索されたドキュメントの組が得られる。その後各ドキュメントには、自然言語処理が施され、特定的には形態素

的、統語論的および論理的な形式に関する処理が施され、最終的には各ドキュメントの各文に対して適当な論理形式が生成される。ユーザによって与えられたクエリが同様に分析され、それに関する対応の論理形式三つ組の集合が得られる。クエリに対する論理形式の組はその後、検索されたドキュメントの各々に関連した論理形式の組と比較され、クエリの組からの論理形式と各ドキュメントの組からの論理形式との一致が確認される。一致を生じないドキュメントはそれ以上は考慮されない。残りの各ドキュメントはその後ヒューリスティックにスコア付けされる。特に、種々のタイプの関係の各々、すなわち論理形式内に現われ得る深層主語、深層目的語および機能語などに、予め規定された重みが割当てられる。このような残りのドキュメントの各々のスコアは、その中にある一致する論理形式の重みの、予め規定された関数である。この関数は、例えば、そのドキュメントに現われる、一意な、一致する三つ組（二重の一致を無視）全てに関連した重みの和でもよい。最後に、保持されたドキュメントが、それらのスコアに基づいて、ユーザの選択にしたがい降順に、典型的にはたとえば5個または10個というような予め規定された少数のグループで、最も高いスコアを有するグループから始まり、順に他のグループという順番でユーザに提示される。

【0019】

本発明は種々のいくつかの処理トポロジで 사용할 ことができる。すなわち、（a）クエリおよびキーワードに基づいたサーチ（ドキュメント検索）の両方がローカルなパーソナルコンピュータ（PC）などの共通のコンピュータによって処理される場合、（b）キーワードに基づいたサーチがたとえばリモートサーバであるリモートコンピュータによって処理され、クエリおよびサーチ結果がたとえばクライアントPCによって処理される場合、または（c）クエリがクライアントPCで作成され、残りの処理がさまざまなリモートサーバに分配される場合、である。さらに、データベースの各ドキュメントを索引付けしてデータベース化する際に前処理して、関連の論理形式を得て、これらの論理形式を記憶しておいて後にアクセスできるようにすることによって、そのドキュメントが後に検索されて自然言語処理を受ける場合には常に実行時間が節約されるようになる。

【0020】

(30)

【詳細な説明】

本発明の教示は添付の図面を参照して以下の詳細な説明を考慮すると容易に理解できる。

【0021】

理解を容易にする目的で、可能な場合には、図面に共通する要素には同一の参照番号を付す。

【0022】

以下の説明を考慮すれば、当業者であれば、サーチエンジンが従来の統計的エンジンであるか否かに関係なく、我々の本発明の教示をほとんど全ての情報検索システムに容易に利用して、そのシステムで使用されるサーチエンジンの精度を高めるようにできることが明らかに認識できるであろう。さらに、我々の発明は、磁気媒体、光媒体（たとえばCD-ROM）または他の媒体に記憶されているデータベースなどのほぼ全てのタイプの大容量記憶装置からテキスト形式の情報を検索する際に、たとえば英語、スペイン語およびドイツ語など、テキスト形式の情報がどの言語であるかにかかわらず、精度を改善するために利用できる。

【0023】

広く言えば、我々の本発明によると、たとえば検索エンジンで使用されているサーチエンジンによって提供されるレコードを、たとえば究極的にはドキュメントを、フィルタリングしランク付けするために自然言語処理を用いることによって、検索エンジンの精度を著しく向上させることができることが我々には認識できた。

【0024】

この点に留意して、図1は我々の発明を利用する情報検索システム5の非常に高いレベルのブロック図を示す。システム5はたとえばキーワードに基づいた統計的検索エンジンである従来の検索エンジン20と、その後続くプロセッサ30とを含む。プロセッサ30は、後に説明するように、我々の発明の自然言語処理技術を利用して、エンジン20によって生成されたドキュメントをフィルタリングして再度ランク付けし、ユーザによって与えられたクエリに対する関連性がこの技術を仕様しない場合よりも高い、順序付けのされた検索されたドキュメン

(31)

トの集合をもたらす。

【0025】

具体的には、動作時にユーザはサーチ用のクエリをシステム5に与える。クエリには、自然言語処理によってその意味論上の内容を最大限に利用し、それによってエンジン20だけにより得られる精度よりもさらに精度を向上させるために、（通常は「リテラル」と呼ばれる）フルテキスト形式が用いられる。システム5はこのクエリをエンジン20およびプロセッサ30の両方に与える。クエリに応答して、エンジン20は、記憶されたドキュメントのデータセット10をサーチし、それから検索されたドキュメントの集合を出力する。このドキュメントの集合（「出力ドキュメント集合」とも呼ばれる。）は、線25で表わされるように、入力としてプロセッサ30に与えられる。後に詳細に説明するように、プロセッサ30内では、集合内のドキュメントの各々に対して、自然言語処理、特に形態素的、統語論的および論理的な形式に関する処理が施され、そのドキュメント内の各文に対する論理形式を生成する。ある文に関するこのような論理形式の各々は、その文内の言語学的な句の中の単語間の意味論的な関係、特に主題および修飾語句構造を符号化したものである。プロセッサ30は同一の態様でクエリを分析し、それに関する対応の論理形式の集合を生成する。そしてプロセッサ30は、クエリに関する形式の集合と、その集合内のドキュメントの各々に関連する論理形式の集合とを比較し、クエリ集合内の論理形式と各ドキュメントに関する論理形式との間に一致があるかどうかを確認する。一致を生まないドキュメントはそれ以上考慮されない。残りの、クエリに関する論理形式と一致する少なくとも1つの論理形式を含むドキュメントの各々は、プロセッサ30によって保持され、ヒューリスティックにスコア付けされる。後に説明するように、異なった各タイプの関係、すなわち、論理形式三つ組に現われ得る深層主語、深層目的語および機能語などに対して、予め規定された重みが割当てられる。このようなドキュメントの各々の合計の重み（すなわちスコア）は、たとえば一致する一意な三つ組、すなわち二重に一致する三つ組を無視したものの全ての重みの和である。最後に、プロセッサ30は、保持されたドキュメントを、それらのスコアに基づいてランク付けして、たとえば5個とか10個とかという予め定められた数の

グループに分けて、スコアの最も高いものからユーザに提示する。

【0026】

システム5が非常に汎用的であり、広範囲な種々のアプリケーションに適合させることができるので、以下の議論を簡単にするために、我々は1つの例を用いて我々の発明の用途を議論することとする。この例は、ワールドワイドウェブからのドキュメントであって、索引付けされたデータセットを形成する英語のドキュメントの、格納されたレコードを検索するために、従来のキーワードに基づいた統計的インターネットサーチエンジンを採用する情報検索システムであろう。このような各レコードは一般に、以下に説明するように、対応のドキュメントに関する予め規定された情報を含む。他のサーチエンジンの場合、レコードがドキュメント自体の全体を含んでもよい。以下の議論では、ドキュメントに関する、そのドキュメントを見出すことができるウェブアドレスを含むある情報を含むレコードを検索する従来のインターネットサーチエンジンに使用する場合を例として我々の発明を扱うが、包括的に言えば、そのエンジンによって検索される究極的な項目とは、ウェブからドキュメントに実際にアクセスするために一般的にはそのアドレスを用いる中間的な処理が採用されるところとしても、実際にはドキュメントである。以下の説明を考慮すると、我々の本発明が、他のいかなる情報検索の適用例の使用にも容易に適合可能であることが当業者には容易に認められるであろう。

【0027】

図2は、インターネットサーチエンジンの例において使用されている、我々の発明の特定の実施例の高いレベルでのブロック図を示す。我々の発明は、主にこの特定の実施例を例として詳細に説明する。図示されるように、システム200はクライアントパーソナルコンピュータ(PC)などのコンピュータシステム300を含み、これは、ネットワーク接続205を介して、ネットワーク210（ここではインターネットであるが、たとえばイントラネットなどの他のこのようなネットワークをこれに代えて用いてもよい）およびネットワーク接続215によってサーバ220に接続される。サーバは典型的にはコンピュータ222を含み、これはインターネットサーチエンジン225をホストし、たとえばALTA

(33)

VISTAサーチエンジン (ALTA VISTAはマサチューセッツ州メイナード (Maynard, Massachusetts) のDigital Equipment Corporationの登録商標である。) が典型であり、大容量データ記憶装置227に接続され、これは典型的には、サーチエンジンによって索引づけられ、インターネット上のワールドワイドウェブによってアクセス可能であるドキュメントレコードのデータセットである。このようなレコードの各々は典型的には、(a) ウェブブラウザによって対応のドキュメントがアクセス可能であるウェブアドレス (通常はユニフォームリソースロケータ—URLと呼ばれる) と、(b) そのドキュメントに現われる予め規定された内容語とを含み、エンジンによっては、そのドキュメント内の、他の内容語に対するこのような各単語の相対的なアドレスを含み、さらに(c) ドキュメントのうち数行だけであるか、またはドキュメントの最初の数行であることが多い概要と、(d) ハイパーテキストマークアップ言語 (HTML) 記述フィールドに提供されるような、ドキュメントの説明とを含む。

【0028】

コンピュータシステム300に配置されたユーザは、このシステムで動作する、たとえば関連のウェブブラウザ (たとえばMicrosoft Corporationから入手可能であり、我々の発明の教示を含むよう適当に変形された「インターネットエクスプローラ」バージョン3.0に基づくもの) を介して、サーバ220、特にそこで動作するサーチエンジン222へのインターネット接続を確立する。さらにユーザは、ここでは線201によって表わされるクエリをブラウザに入力し、ブラウザはシステム300を介して、サーバ220へのインターネット接続によって、サーチエンジン225にクエリを送信する。するとサーチエンジンはデータセット227に記憶されたドキュメントレコードに対してクエリを処理し、エンジンがクエリに関連すると判断したドキュメントに対する検索レコードの集合を生成する。エンジン225がどのようにしてドキュメントを索引付けしドキュメントレコードを形成してデータ記憶装置227に記憶するか、および記憶されたこのようなドキュメントを選択するためにエンジンが実際にはどのような分析を行なうかはいずれも本発明と無関係であるため、これらの局面はいずれもこれ以上詳細には説明しない。クエリに応答して、エンジン225が、検索されたドク

(34)

ュメントレコードの集合をインターネット接続を介してウェブブラウザ420に返す、といえは十分である。ブラウザ420は、エンジン225がドキュメントを検索すると同時に、かつ／またはその後、クエリを分析して、その、論理形式三つ組の対応の集合を生成する。サーチエンジンがそのサーチを完了し、ドキュメントレコードの集合を取出し、その集合をブラウザに与えると、対応のドキュメント（すなわち出力ドキュメントの集合を形成するもの）自体が関連のウェブサーバからブラウザによってアクセスされる（これに関連したデータセットは全体として保存されたドキュメントの「リポジトリ」を形成する。このようなリポジトリは、たとえば独立したCD-ROMベースのデータ検索アプリケーションなどにおけるもののよう、スタンドアローンのデータセットであってもよい）。するとブラウザはアクセスされたドキュメント（すなわち出力ドキュメントの内のもの）の各々を分析し、このようなドキュメントの各々に関する、論理形式三つ組の、対応する集合を形成する。その後、後に詳細に説明するように、ブラウザ420は、クエリと、検索されたドキュメントとの間の論理形式三つ組の一致照合に基づいて、このような一致を有する各ドキュメントをスコア付けし、それらのドキュメントを、線203によって表わされるようにランクの降順に、ブラウザを通じたユーザの選択にしたがって、典型的には最も高いランクを有する少数の予め規定されたドキュメントのグループとして、大きなスコアのものからランク付けされてユーザに提示され、さらにこの後、ユーザがこのように提示されたドキュメントの十分な数を確認するまで、次のグループが続き、以下同様である。図2はリモートサーバからドキュメントレコードおよびドキュメントを獲得するためにネットワーク接続を例示的に利用するものとして我々の発明を示すが、我々の発明はこのようには限定されない。図9Aに関連して以下に詳細に説明するように、検索アプリケーションおよび我々の発明が共通のコンピュータ、すなわちローカルPC上で実行され、そこにたとえばCD-ROMまたは他の適切な媒体に記憶された付随するデータセットが配置されアクセス可能であれば、このようなネットワーク接続は必要ない。

【0029】

図3は、図2に示されるコンピュータシステム300のブロック図を示し、こ

(35)

のコンピュータシステム300は本発明の教示を取入れたものである。

【0030】

示されるように、例えばクライアントパーソナルコンピュータであるこのシステムは、全て従来からバス370によって相互接続されている入力インターフェイス (INPUT I/F) 330と、プロセッサ340と、通信インターフェイス (COMM I/F) 350と、メモリ375と、出力インターフェイス (OUTPUT I/F) 360とを含む。メモリ375は例えばランダムアクセスメモリ (RAM) およびハードディスク記憶装置である種々の様式 (これらは全て簡略化のために特に示してはいない) を一般的に含むが、オペレーティングシステム (O/S) 378と、アプリケーションプログラム400とを記憶する。我々の発明の教示を実装するソフトウェアは典型的にはアプリケーションプログラム400に組込まれ、この実施例では特にウェブブラウザ (図4に示される) に組込まれる。このオペレーティングシステムは、ウィンドウズNTオペレーティングシステム (ワシントン州レッドモンド (Redmond, Washington) のMicrosoft Corporation (登録商標「ウィンドウズNT」も所有する) から現在入手可能である。) など、どのような従来のオペレーティングシステムによって実装されてもよい。O/S 378の構成要素であるプロセスは発明とは無関係であるため、各部分については説明しない。しかしながら、ブラウザ、従って我々の発明のソフトウェアを、オペレーティングシステム自体の中に組込むこともできる。しかし、例示および簡略化の目的のために我々は、ブラウザがオペレーティングシステムから分離可能であり、アプリケーションプログラム400内にあると仮定する。アプリケーションプログラム400はO/S 378の制御下で実行される。ウェブブラウザを含む実行アプリケーションプログラムの各々に対して、1つまたは2つ以上の別個のタスクのインスタンスが、ユーザが特定した各コマンド、典型的にはメニューとか、ツールバー内のアイコンなど選択可能なコマンドが利用可能な場合にユーザが入力でバイス390を適切に操作することによって対話的に入力されるコマンドに応答して、ユーザによって呼出され、付随する情報がディスプレイ380に提示される。

【0031】

図 3 に示されるように、入来する情報は、例えば 2 つの外部ソースから生ずる。すなわち、たとえばインターネットおよび／またはイントラネットなどのネットワーク化された他の設備（いずれも全体的に図 2 にネットワーク 210 として示される）から、ネットワーク接続 205 を介して通信インターフェイス 350（図 3 に示される）へ達するネットワーク空供給される情報、または経路 310 を介して専用の入力ソースから入力インターフェイス 330 へ達するものである。専用の入力とは、たとえばローカルであると、リモートであると、または他の入力ソースであるにもかかわらず、外部のデータセットなどさまざまなソースから生ずる。入力インターフェイス 330 は経路 310 に接続され、入力情報の、種々の専用のソースの各々をコンピュータシステム 300 に物理的に接続してインターフェイスするために必要である対応の電気接続を提供する適当な回路構成を含む。アプリケーションプログラム 400 は、オペレーティングシステムの制御下で、ネットワーク接続 205 を介してリモートウェブサーバなどの外部ソースと、または経路 310 を介して専用のソースなどと、コマンドおよびデータを交換し、プログラムの実行時に典型的にはユーザによってリクエストされる情報の送受信を行なう。

【0032】

入力インターフェイス 330 は、リード 395 によって、キーボードおよびマウスなどのユーザ入力デバイス 390 をコンピュータシステム 300 に電氣的に接続し、インターフェイスする。従来のカラーモニタなどのディスプレイ 380 および従来のレーザプリンタなどのプリンタ 385 は、それぞれリード 363 および 367 によって出力インターフェイス 360 に接続される。出力インターフェイスはディスプレイおよびプリンタをコンピュータシステムに電氣的に接続してインターフェイスさせるために不可欠な回路構成を提供する。実行中のアプリケーションからのハードコピー出力情報はプリンタ 385 によってユーザに与えられる。特に、ディスプレイ、プリンタおよび入力デバイス 390（特定のにはマウスおよびキーボード）を適切に操作することにより、システム 300 に配置されているユーザは、たとえば、インターネットを介して、そこからさらにアクセス可能なサーチエンジンを含む、膨大な数のリモートウェブサーバのうちのい

(37)

ずれかと画面を使って通信し、ローカルに表示および印刷するためにそこからドキュメントなどの情報を引出すことができる。

【0033】

本発明の実現に必要なもの以外の、コンピュータシステム300の特定のハードウェア構成要素およびメモリ375に記憶されたソフトウェアのそれぞれの局面は従来からのものであり周知であるため、これ以上詳細には説明しない。

【0034】

図4は、図3に示されるコンピュータ300内で実行されるアプリケーションプログラム400の非常に高いレベルのブロック図を示す。これらのプログラムは本発明に関連する範囲では、図4に示されるようにウェブブラウザ420を含み、このウェブブラウザ420は、我々の本発明を実現するための検索プロセス600（図6Aおよび図6Bに関連して後に詳細に説明される）を含む。ウェブブラウザと、ALTA VISTAサーチエンジンなどのユーザが選択した統計的サーチエンジンとの間にインターネット接続が確立されているものと想定すると、ユーザは、図4に示される線422で表わされるように、プロセス600にフルテキスト（「リテラル」）サーチ用のクエリを与える。このプロセスは、線426で表わされるように、ウェブブラウザを介してサーチエンジンにクエリを転送する。さらに、特に示してはいないが、プロセス600はさらにクエリを内部で分析し、その対応の論理形式三つ組を生成し、これらは後にコンピュータ300内にローカルに記憶される。クエリに応答して、サーチエンジンは線432で表わされるように統計的に検索されたドキュメントレコードの集合をプロセス600に与える。これらのレコードの各々は、上述のとおり、のドキュメントをアクセスすることができるウェブアドレス、特定的にはURLと、さらに、そのドキュメントがあるリモートウェブサーバによって要求される、そのドキュメントを含むコンピュータファイルをインターネットを介してダウンロードするのに十分な、適切なコマンドを含む。プロセス600がレコード全てを受信すると、このプロセスは、ウェブブラウザ420を介して、かつ線436によって表わされるように適切なコマンドを送信し、レコードによって特定された全てのドキュメント（すなわち出力ドキュメントの集合を形成するもの）にアクセスしてそれ

らをダウンロードしようとする。そしてこれらのドキュメントは、それらに対応するウェブサーバから順次アクセスされ、線442で表わされるように、ウェブブラウザ420に、特定的にはプロセス600にダウンロードされる。これらのドキュメントがダウンロードされると、プロセス600はこのようなドキュメントの各々を分析して、それに関する対応の論理形式三つ組を生成してローカルに記憶する。その後、各ドキュメントに関する論理形式三つ組に対して、クエリに関する論理形式三つ組を比較することにより、プロセス600は少なくとも1つの一致する論理形式三つ組を含む各ドキュメントをスコア付けし、それらのスコアに基づいてこれらの特定のドキュメントをランク付け、最後に、線446によって表わされるように、ドキュメントのスコアの降順に、グループ毎にこれらの特定のドキュメントをユーザに提示するようにブラウザ400に対して指示する。ブラウザ400はディスプレイ380（図3参照）のスクリーン上に適切な選択ボタンを作成し、これによりユーザは、ユーザのマウスでそれを適切に「クリック」することによって選択を行ない、所望にしたがって、後続するドキュメントのグループの各々を表示することができる。

【0035】

意味論上の情報を判断し、保存し、符号化する際の論理形式の有用性を十分に評価するために、この時点で、我々の発明を実現する処理の説明から離れ、関連する範囲で本発明に用いられる論理形式および論理形式三つ組を例示して説明し、それらが生成される態様を簡単に説明する。

【0036】

大まかに言って、論理形式は、任意のサイズを有するテキストを表わす単語がラベル付けされた関係によってリンクされる、有向非循環グラフである。論理形式は、句内の重要な単語、この単語には上位語および／またはその類義語を含めることもあり得るが、これら単語間の意味論的な関係を描く。図5Aから図5Dを参照して説明され例示されるように、論理形式は種々の多くの形態のうちいずれかをとることができ、たとえば論理形式三つ組のリストなどの、論理形式グラフまたは任意のそのサブグラフの形式をとることができる。たとえばこれら三つ組の各々は「語－関係子－語」という形式を持つ。この実施例では、本発明は論

理形式三つ組を生成して比較するが、本発明は、単語間の意味論的な関係を描くことができるものであれば、上述したような他のいかなる形式を容易に利用することができる。

【0037】

論理形式三つ組およびそれらの構造は順により複雑になるような一連の文を例とすることによって最もよく理解できるため、まず図5Aを参照する。この図は、例示的な入力文字列、特定的には「The octopus has three hearts.」という文に関する論理形式グラフ515と論理形式三つ組525とを示す。

【0038】

一般に、入力文字列、たとえば入力文字列510に対する論理形式三つ組を生成するために、この文字列はまずパーズングされてその構成要素の単語に分解される。その後、このような各単語に対して、予め格納された辞書にある、予め定義されたレコード（サーチエンジンによって採用されるドキュメントレコードと混乱してはならない。）を用いて、これらの構成要素の単語に対応のレコードが、予め規定された文法規則によって組み合わせられて大きな構造または構文になり、さらにそれらは予め定められた文法規則によって再度組み合わせられて、構文解析木などのさらに大きな構造を形成する。その後論理形式グラフが解析木から構築される。特定の規則が特定の構成要素の集合に適用可能であるか否かは、部分的には、単語レコードに、ある対応の属性およびそれらの値が存在するか否かによって支配される。そしてこの論理形式グラフは、一連の論理形式三つ組に変換される。例えば、我々の発明はおよそ165,000個のヘッド単語のエントリを有する辞書を使用する。この辞書は入力文字列に関する解析木が構築できるように、入力文字列内の単語に固有な統語論上および意味論上の特性を規定する、前置詞、接続詞、動詞、名詞、オペレータおよび数量詞などの、さまざまなクラスの単語を含む。明らかに、論理形式（またはその問題に関しては、意味論上の関係を描くことができる、論理形式内の論理形式三つ組または論理形式グラフ）を、後に対応のドキュメントが検索されたときに計算するのではなく、対応のドキュメントが索引付けされている間に、予め計算してたとえばそのドキュメントに関するレコード内に格納して、後のアクセスに使用するようにできる。図

(40)

10から図13Bに関連して後に説明する我々の発明の別の実施例に見られるように、このように予め計算して格納しておく、自然言語処理の量が劇的に低減するという効果があり、したがって、我々の発明に従って、検索されたドキュメントを扱うために必要な関連の実行時間が短くなる。

【0039】

特に、図5Aに示される文510などの入力文字列は、まず、その構成要素の単語の各々に関して辞書内にある予め規定されたレコードを用いて形態素分析され、それに関するいわゆる「語幹」（または「基体」）形式を生成する。語幹は、たとえば動詞の時制および単数一複数といった名詞の変形などの種々の単語の形を、パーサが使用できるような共通の形態素的形式に正規化するために用いられる。一旦語幹形式が作成されると、文法規則および構成要素の単語のレコード内にある属性を用いて、入力文字列がパーサによって構文解析され、それに関する構文解析木が生成される。この木は入力文字列の構造を示し、具体的には、たとえば入力文字列内の「The octopus」という名詞句のような各単語または句と、たとえば名詞句に対するNPのような、対応の文法的機能のカテゴリと、その中の、構文的に関連した各単語または句へのリンクとを表わす。例示的な文510については、関連の構文解析木は以下のとおりであろう。

【0040】

【表1】

| | | |
|------|--------------|------------------------|
| DECL | ---NP | --- DETP-ADJ* "The" |
| | | --- NOUN* "octopus" |
| | ---VERB* has | |
| | ---NP | --- QUANP-ADJ* "three" |
| | | --- NOUN* "hearts" |
| | ---CHAR ". " | |

【0041】

表1 -- 「The octopus has three hearts.」に関する構文解析木

(41)

木の上部左側にある開始ノードは、パーシングされる入力文字列のタイプを定義する。文のタイプには、平叙文に関する「DECL」（上の例）と、命令文に関する「IMPR」と、疑問文に関する「QUES」とが含まれる。開始ノードの右下垂直に表示されるのは、第1レベルの構文である。この構文は、典型的には主動詞（この例では「has」という単語）である、星印によって示されたヘッドノードと、（この例では「The octopus」という名詞句である）前置修飾語句と、その後にくる（「three hearts」という名詞句である）修飾語句とを有する。木の葉の各々は辞書に含まれる単語または句読点を含む。ここでは、ラベルとして「NP」は名詞句を示し、「CHAR」は句読点を示す。

【0042】

そして構文解析木は、異なった組の規則を用いてさらに処理され、入力文字列 510 に関するグラフ 515 などの論理形式グラフが生成される。論理形式グラフを作成するプロセスは、入力文字列の構文解析から下層の構造を抽出することを含み、論理形式グラフは、互いの間に意味関係と、その関係の機能的性質を有すると定義された複数の単語を含む。種々の意味関係の分類に使用される「深層」の格 (Case) すなわち機能的役割は、以下を含む。

【0043】

- D s u b — 深層主語
- D i n d — 深層間接目的語
- D o b j — 深層目的語
- D n o m — 深層述語主格
- D c m p — 深層目的格補語

表 2

入力文字列内の意味関係全てを特定するために、その文字列に関する構文解析木の各ノードが検査される。上記の関係に加えて、たとえば下記の他の意味的役割が使用される。

【0044】

- P R E D — 述部
- P T C L — 2部構成の動詞における不変化詞

(42)

| | |
|-----------|---------------------|
| O p s | ――機能語、たとえば数字 |
| N a d j | ――名詞を修飾する形容詞 |
| D a d j | ――叙述形容詞 |
| P R O P S | ――節である、他には特定されない修飾部 |
| M O D S | ――節でない、他には特定されない修飾部 |

表 3

同様に追加の意味的ラベルが下記のように規定される。

【 0 0 4 5 】

| | |
|-----------|------|
| T m e A t | ――時刻 |
| L o c A t | ――場所 |

表 4

いずれにせよ、入力文字列 5 1 0 に関するこのような分析の結果は論理形式グラフ 5 1 5 である。入力文字列中の単語で互いの間に（たとえば「Octopus」と「Have」との間の）意味関係が認められる単語は、互いにリンク付けられ、それらの間の関係はリンク付け属性（たとえば D s u b）として特定されて示されている。このグラフは、入力文字列 5 1 0 に関するグラフ 5 1 5 に代表されるように、各入力文字列に関する主題および修飾語句の構造を捕らえている。とりわけ、論理形式分析は、前置詞および冠詞などの機能的単語を、グラフ内に示された特徴または構造上の関係にマッピングする。論理形式分析はさらに前方照応を解決、すなわち、たとえば代名詞と、同一指示名詞句との間の正しい先行関係を規定し、さらに省略に関する適切な機能的関係を検出して示す。論理形式分析時には、曖昧さおよび／または他の言語的特異性に対処するためにさらに処理が施されることもあり得る。そして、対応の論理形式三つ組が従来の態様で論理形式グラフから読出され、組として記憶される。各三つ組は、グラフに示されるように、互いの間の意味関係によってリンク付けられた 2 つのノード単語を含む。例としての入力文字列 5 1 0 に関しては、論理形式三つ組 5 2 5 が処理グラフ 5 1 5 から結果として得られる。ここでは、論理形式三つ組 5 2 5 は、入力文字列 5 1 0 に固有の意味論的な情報を全体として伝える 3 つの別個の三つ組を含む。

【 0 0 4 6 】

同様に、図 5 B から図 5 D に示されるように、入力文字列 5 3 0、5 5 0 および 5 7 0、すなわち例示の文「The octopus has three hearts and two lungs.」、「The octopus has three hearts and it can swim.」、および「I like shark fin soup bowls.」に対しては、論理形式グラフ 5 3.5、5 5 5 および 5 7 5 ならびに論理形式三つ組 5 4 0、5 6 0 および 5 8 0 がそれぞれ結果として得られる。

【0047】

論理形式三つ組が論理形式グラフから生成される従来の「グラフウォーク」を含む従来の態様のものとは別に、論理形式三つ組の全てを正しく得るために必要な追加の自然言語処理が必要な、3つの論理形式構造がある。例示の文「The octopus has three hearts and two lungs.」、すなわち入力文字列 5 3 0 におけるような等位語の場合、単語と、その意味関係と、等位された構成要素の値の各々々々に対する論理形式三つ組が生成される。「特殊な」グラフウォークによると、図 5 4 0 には 2つの論理形式三つ組「have-Dobj-heart」および「have-Dobj-lung」があることがわかる。従来のグラフウォークのみを用いると、1つの論理形式三つ組「have-Dobj-and」しか得られなかったであろう。同様に、指示語 (Refs) を有する構成要素の場合、例示の文「The octopus has three hearts and it can swim.」、すなわち入力文字列 5 5 0 の場合の様に、従来のグラフウォークによって生成された三つ組の他に、単語と、その意味関係と、Refs属性の値の各々々々に対する論理形式三つ組を生成する。この特殊なグラフウォークによると、従来の論理形式三つ組「swim-Dsub-it」の他に論理形式三つ組「swim-Dsub-octopus」が三つ組 5 6 0 に見出される。最後に、例示の文「I like shark fin soup bowls.」、すなわち入力文字列 5 7 0 におけるように、名詞の修飾語で構成される場合、名詞の複合語の可能な内部構造を表わすためにさらなる論理形式三つ組が生成される。従来のグラフウォークでは、可能な内部構造[[shark] [fin] [soup] bowl]を反映する論理形式三つ組「bowl-Mods-shark」、「bowl-Mods-fin」および「bowl-Mods-soup」が生成される。特殊なグラフウォークの場合、下記の可能な内部構造[[shark fin] [soup] bowl]、[[shark] [fin soup] bowl]および[[shark [fin] soup] bowl]を表わすためにそれぞれ追加の論理形式三つ組

(44)

「fin-Mods-shark」、「soup-Mods-fin」、および「soup-Mods-shark」が生成される。

【0048】

形態素的、統語論的および論理形式処理の特定の詳細は本発明とは関係ないため、さらに詳細な説明は省略することとする。しかしながら、これに関するさらなる詳細は、1996年6月28日に出願され、連続番号第08/674,610号が付与された「構文解析木から意味論的論理形式を計算するための方法およびシステム（“Method and System for Computing Semantic Logical Forms from Syntax Trees”）」と題された同時係属中の米国特許出願と、特に1997年3月7日に出願された、連続番号_____が付与された「テキストの意味表現を利用する情報検索（“Information Retrieval Utilizing Semantic Representation of Text”）」とを参照されたい。これらはいずれも本願の譲受人に譲渡されており、引用によってここに援用される。

【0049】

この論理形式の概要およびそれらの構造を念頭に、我々の本発明を実現する処理の議論に戻ることとする。

【0050】

図2、図3、図4に示される我々の発明の特定の実施例に使用されている、我々の発明の検索プロセス600のフローチャートは、図6Aおよび図6Bにまとめて示され、これらの図面の正確な配列が図6に示される。破線で描かれたブロック225に示される動作以外の、これらの図面に示される残りの動作は、たとえばクライアントPC300（図2および図3参照）であるコンピュータシステムによって行なわれ、具体的にはウェブブラウザ420内で行なわれる。理解を容易にするために、以下の説明を読みながら図2、図3、図6Aおよび図6Bを同時に参照すべきである。

【0051】

プロセス600に入ると、実行処理はまずブロック605に進む。このブロックは、実行されると、フルテキスト（リテラル）のクエリをウェブブラウザ420に入力することをユーザーに促す。クエリは単一の質問（たとえば「Are ther

(45)

e any air-conditioned hotels in Bali?」)でも、単一の文(たとえば「Give me contact information for all fireworks held in Seattle during the month of July.」)でも、文の一部(たとえば「Clothes in Ecuador」)の形式であってもよい。この質問が得られると、実行処理は経路607を介してブロック610に、および経路643を介して経路645に分岐する。ブロック645が実行されると、それはNLPルーチン700を呼出し、クエリを分析し、その対応の論理形式三つ組の集合を構築してローカルに記憶する。ブロック610が実行されると、それは、破線615で表わされるように、インターネット接続によって、フルテキストのクエリを、ウェブブラウザ620から、サーバ220に置かれたエンジン225などのリモートサーチエンジンに送信する。この時点で、ブロック625がサーチエンジンによって実行され、クエリに応答してドキュメントレコードの集合を取出す。この集合が形成されると、破線630で表わされるように、その集合はリモートサーバによってコンピュータシステム300に再送信され、特に、そこで実行されているウェブブラウザ420に戻される。その後、ブロック635が実行され、レコードの集合を受信し、後に、各レコードに対して、そのレコードからURLを抽出し、そのURLでウェブサイトにアクセスし、さらにはそのレコードに対応するドキュメントを含む関連ファイルをそこからダウンロードする。全てのドキュメントがダウンロードされると、ブロック640が実行される。このような各ドキュメントに対して、このブロックはまず、そのドキュメントに関連するHTMLタグ内にある全てのテキストを含む、全てのテキストをそのドキュメントから抽出する。その後、一度に一つの文に対して行なわれる自然言語処理を容易にするために、各ドキュメントに関するテキストが、従来の一文切出し処理によって各文章(または質問)がファイル内の別々の行を占めるようなテキストファイルに切出される。その後、ブロック640は、そのドキュメントのテキストの各行に対して、NLPルーチン700(これは図7に関連して後に詳細に説明される)を繰返して呼出し、これらのドキュメントの各々を分析し、そのドキュメントのテキストの各行に関する対応の論理形式三つ組の集合を構築してローカルに記憶する。ブロック645における動作は、ブロック610、635および640におけるものと基本的に並行して行なわれ

(46)

るものとして説明したが、前者のブロックの動作は、実際の実装での条件に基づいて、ブロック610、635および640の動作と順次に、それらの前または後のいずれかに行なってもよい。これに代えて、図10から図13Bに関連して後に説明する我々の発明の別の実施例の場合のように、各ドキュメントに関する論理形式三つ組を予め計算して記憶しておき、後のドキュメントの検索時に使用してもよく、この場合にはこれらの三つ組はドキュメントの検索時には計算されずに単にアクセスされるだけである。この場合、これら三つ組は、何らかの態様で、その格納されたドキュメントのプロパティ（属性）として、または、たとえばそのドキュメントに関するレコードまたはそのドキュメントを含むデータセットのいずれかに別のエントリとして、記憶されるであろう。

【0052】

いずれにせよ、図6Aおよび図6Bに示されるプロセス600に戻り、論理形式三つ組の集合がクエリおよび出力ドキュメントの集合内の検索されたドキュメントの各々の双方に対して構築され完全に記憶されると、ブロック650が実行される。このブロックは、クエリの論理形式三つ組の各々と、検索されたドキュメントの各々に関する論理形式三つ組の各々とを比較して、クエリのいずれかの三つ組と、ドキュメントのいずれかの三つ組のいずれかとの一致を突き止める。例としての一致の形式は、これらの三つ組の間での、ノード単語と、関係子タイプとの双方の点においてこれら2つの三つ組間で同一の一致が見られることと定義される。特に、例示の1対の論理形式三つ組、すなわち「語1a－関係子1－語2iおよび語1d－関係子2－語2b」の場合、ノード単語の語1aおよび語1bが同一であり、ノード単語の語2aおよび語2bが同一であり、関係子1および関係子2が同じである場合にのみ一致が起こる。1つの三つ組の3つの要素の全てが別の三つ組の対応する要素と同一に一致しないならば、これらの2つの三つ組は一致しない。ブロック650が完了すると、ブロック655が実行され、一致する三つ組が得られない、すなわちクエリの三つ組と一致する三つ組がない、検索された全てのドキュメントが破棄される。その後、ブロック660が実行される。ブロック660により、一致する三つ組の関係子のタイプ、およびこれらのドキュメントの各々に関して存在するそれらの重みに基づいて、残りのド

(47)

キュメント全てにスコアが割当てられる。特に、論理形式三つ組に生じ得る異なったタイプの関係子の各々に、図8Aの表800に示されるもののような対応の重みが割当てられる。たとえば、図示されるように、例示の関係子Dobj、Dsub、OpsおよびNadjにはそれぞれ、100、75、10および10という、予め定められた静的な数の重みを割当てることができる。重みはクエリとドキュメントとの正しい意味上の一致を示す上で、その関係子に帰すると考えられる相対的な重要性を反映する。これらの重みの実際の数値は一般に、経験に基づいて定義される。後に図8Bに関連して詳細に説明するように、残りのドキュメントの各々に関して、そのスコアは予め定義された関数であり、ここでは例として、その一意に一致する三つ組（二重に一致する三つ組は全て無視する。）の重みの数値の和である。こうしてドキュメントが一旦重み付けされると、ブロック665が実行されて、スコアの降順に、ドキュメントをランク付ける。最後に、ブロック670が実行されて、典型的には、最も高いスコアを示す予め規定された小さなグループのドキュメント、典型的には5個から10個のドキュメントをランク順に表示する。その後、ユーザーは、たとえばウェブブラウザ420によって表示された対応のボタン上でユーザーのマウスを適当に「クリック」することにより、コンピュータシステム（クライアントPC）300に、ランク付けされたドキュメントの次のグループを表示させ、以下同様に、ユーザーがランク付けされたドキュメント全てを順次十分に検討するまでこれを続け、そこでプロセス600を完了する。

【0053】

図7は、NLPルーチン700のフローチャートを示す。このルーチンは、1行の入力テキストを与えられ、そのテキストがクエリ、ドキュメント内の文、またはテキストの一部のいずれの場合にも、それに関する対応の論理形式三つ組を構築する。

【0054】

特に、ルーチン700に入ると同時に、ブロック710はまず入力テキストの行を処理し、図5Aに示される例示のグラフ515などの論理形式グラフを生成する。この処理は、構文解析木を生成する形態素的および統語論上の処理を含み

(48)

、この構文解析木から後に論理形式グラフが算出される。その後、図7に示されるように、ブロック720が実行されて、グラフから対応の論理形式三つ組の集合を抽出する（読出す）。これが行なわれると、ブロック730が実行されて、このような論理形式三つ組の各々を、別個で区別された、フォーマット化されたテキスト文字列として生成する。最後に、ブロック740が実行されて、データセット（すなわちデータベース）に入力テキストの行を記憶し、フォーマット化された一連のテキスト文字列として、その行に関する論理形式三つ組の集合を記憶する。この集合が完全に記憶されると、実行処理はブロック700を出る。これに代えて、論理形式三つ組の代わりに、たとえば論理形式グラフのような、異なった表現が論理形式に関連付けられて我々の発明に関し用いられる場合には、その特定の形式をフォーマット化された文字列として生成するようにブロック720および730を容易に変更でき、データセットへの論理形式三つ組の代わりにその形式を記憶するようにブロック740を変更できる。

【0055】

論理形式三つ組の一致を比較しそれに重み付けをし、さらには対応のドキュメントをランク付ける、我々の発明の態様を十分に認識するために、図8Bを参照する。この図は、我々の発明の教示に従う論理形式三つ組の比較、ドキュメントの記憶、ランク付けおよび選択処理を図で示し、これらは、例示のクエリおよび検索された3つのドキュメントの例示の組に関する、図6Aおよび図6Bに全て示されるブロック650、660、665および670において行なわれる。例示の目的で、ユーザーがフルテキストのクエリ810を我々の発明の検索システムに与えたと仮定し、このクエリが「How many hearts does an octopus have?」というものであるとする。さらに、このクエリに応答して、統計的サーチエンジンによって最終的に3つのドキュメント820が検索されたものとする。これらのドキュメントのうち、第1のドキュメント（ドキュメント1と記す。）はartichoke heartsおよびoctopusを含むレシピである。第2のドキュメント（ドキュメント2と記す。）はoctopi（タコ一般）に関する論文である。第3のドキュメント（ドキュメント3と記す。）はdeer（鹿）に関する論文である。これらの3つのドキュメントおよびクエリはそれらの構成要素の論理形式三つ組に変換さ

(49)

れ、それらに関する処理は包括的に「NLP」（自然言語処理）によって表わされる。結果として得られる、クエリとドキュメント1、ドキュメント2およびドキュメント3とに関する論理形式三つ組は、それぞれブロック830、840、850および860に与えられる。

【0056】

これらの三つ組が一旦こうして定義されると、破線845、855および865で表わされるように、クエリに関する論理形式三つ組がそれぞれドキュメント1、ドキュメント2およびドキュメント3に関する論理形式三つ組と順次比較され、いずれかのドキュメントが、クエリのいずれかの論理形式三つ組と一致する三つ組を含むか否かが確認される。ドキュメント1の場合のように、このような一致する三つ組を含まないドキュメントは破棄され、さらに考慮されない。一方、ドキュメント2およびドキュメント3は一致する三つ組を含む。特に、ドキュメント2はこのような三つ組を3つ含む。すなわち、これらはたとえば1つの文に関連する「HAVE-Dsub-OCTOPUS」および「HAVE-Dsub-HEART」と、たとえば別の文に関連する「HAVE-Dsub-OCTOPUS」である（これらの文は特定的には示さない）。これらの三つ組のうち2つは同一であり、すなわちそれは「HAVE-Dsub-OCTOPUS」である。ドキュメントに関するスコアは例えば、そのドキュメント内の、一意に一致する三つ組全ての重みの数値の和である。全てのドキュメントに関し、二重に一致する三つ組は全て無視される。三つ組に生じ得る異なったタイプの関係子の相対的な重みの例示的なランク付けは、最も大きな重みから小さな重みの順に、最初が動詞－目的語の組合せ（Dobj）、動詞－主語の組合せ（Dsub）、前置詞および機能語（たとえば（Ops）、および最後には修飾語句（たとえばNadj）である。このような重み付け方式を、図8Aに示される例示の三つ組重み付け表800に示す。この図を簡単にするために、表800は論理形式三つ組に生じ得る種々の関係子全てを含むのではなく、図8Bに示される三つ組に関連するものだけを示す。この測定基準により、各ドキュメントのうちそのスコアに寄与する特定の三つ組にチェック（「レ」）マークを付与してある。もちろん、ドキュメントにスコア付けするための予め規定される測定基準として我々が選択したもの以外を用いてもよく、たとえば、ドキュメントの選択性（区別）を高めるた

(50)

めに重みを加算するのではなく乗算すること、または重みを別の態様で加算すること、たとえば同じタイプの複数の一致を含めることおよび／または上述以外の他の三つ組の重みを除くことなどであってもよい。さらに、任意のドキュメントに関して、スコアはある態様で下記のことを考慮に入れるであろう。すなわち、そのドキュメントにおける三つ組自体のノード単語、そのドキュメント内のこれらのノード単語の頻度または意味的な内容、そのドキュメント内の特定のノードの単語の頻度または意味的な内容、またはそのドキュメント内の特定の論理形式（またはそのパラフレーズ） および／または特定の論理形式三つ組全体としての頻度、ならびにそのドキュメントの長さである。

【0057】

したがって、我々が選択したスコア付けの測定基準および図8Aの表800に明記される重みを考慮すると、ドキュメント2のスコアは175であり、これはドキュメント内の、ブロック850に示される第1の文に関連した最初の2つの三つ組の重み、すなわち100および75を組合せることにより形成される。このドキュメント内の、その第2の文に関連し、このブロックに記載されている第3の三つ組であって、ドキュメント内に存在する他の三つ組の1つと既に一致しているものは無視される。同様に、ドキュメント3に関するスコアは100であり、これは、ブロック860に記載されているようにこの特定のドキュメント内で唯一の一致する三つ組に関する重み、すなわちここでは100により形成される。スコアに基づいて、ドキュメント2がドキュメント3より前にランク付けされ、これらのドキュメントはその順番でユーザーに提示される。ここでは起こらなかったが、いずれか2つのドキュメントが同じスコアを有する場合、これらのドキュメントは従来の統計的サーチエンジンによって提供されるのと同じ順番でランク付けされ、その順番でユーザーに提示される。

【0058】

明らかに、当業者であれば、我々の本発明を実現するために使用される処理のさまざまな部分が、単一のコンピュータ内に存在しても、全体として情報検索システムを形成する種々のコンピュータに分散してもよいことが容易に理解されるであろう。この点に関し、図9Aから図9Cはそれぞれ、我々の本発明の教示を

採用した情報検索システムの異なった3つの実施例を示す。

【0059】

このような代替的な実施例の1つが図9Aに示され、ここでは全ての処理がPCなどの単一のローカルコンピュータ910によって行なわれる。この場合、コンピュータ910はサーチエンジンをホストし、そのエンジンによって、入力ドキュメントを索引付け、ユーザーによって与えられたフルテキストのクエリに回答して（CD-ROMまたは他の記憶媒体などによってそこにローカルに置かれるか、またはそのコンピュータにアクセス可能である）データセットをサーチし、出力ドキュメント集合を形成する、検索されたドキュメントの集合を最終的に生成する。このコンピュータはさらに我々の発明の処理をホストし、クエリおよびこのような各ドキュメントの両方を分析して、対応の論理形式三つ組の集合を生成し、その後三つ組の集合を比較し、上述の態様でドキュメントをスコア付けてランク付け、最後に、たとえばそこに配置されている、またはそこにアクセス可能なローカルユーザーに結果を提示する。

【0060】

別の代替的な実施例が図9Bに示され、この図9Bは図2に示される特定の状況を包含するものであって、ここでは検索システムはリモートサーバにネットワーク接続されたクライアントPCで形成される。ここでは、クライアントPC920はネットワーク接続925によってリモートコンピュータ（サーバ）930に接続される。クライアントPC920にいるユーザーはフルテキストのクエリを入力し、PCはこれをネットワーク接続を介してリモートサーバに送信する。クライアントPCはさらにクエリを分析して、その対応の論理形式三つ組の集合を生成する。サーバは、たとえば従来の統計的サーチエンジンをホストし、その結果このクエリに回答して、統計的検索を行ない、ドキュメントレコードの集合を生成する。そしてサーバはレコードの集合を戻し、最終的に、クライアントの命令によって、またはサーチエンジンまたは関連ソフトウェアの能力に基づいて自律的に、出力ドキュメントの集合内にある各ドキュメントをクライアントPCに戻す。そしてクライアントPCは出力ドキュメントの集合内の、受信した対応のドキュメントの各々を分析し、それに関する論理形式三つ組の集合を生成する。

(52)

。クライアントPCは次いで、三つ組の集合を適切に比較し、上述の態様でドキュメントを選択し、スコア付けし、ランク付けし、最後に結果をローカルユーザーに提示して、その処理を完了する。

【0061】

さらなる実施例を図9Cに示す。この実施例は図9Bと同じ物理的ハードウェアおよびネットワーク接続を用いるが、クライアントPC920がローカルユーザーからのフルテキストクエリの依頼を受入れ、そのクエリの依頼をネットワーク接続925を介してリモートコンピュータ（サーバ）930へ転送する。このサーバは単に従来のサーチエンジンをホストするのではなく、本発明に従う自然言語処理を提供する。この場合、クライアントPCではなくサーバがクエリを適切に分析してそのための論理形式三つ組の対応の集合を生じるであろう。サーバはまた必要であれば出力ドキュメント集合内の検索された各ドキュメントをダウンロードし、次にこのような各ドキュメントを分析してそのための論理形式三つ組の対応の集合を生成するであろう。その後、サーバはクエリのための三つ組の集合とドキュメントとを適切に比較し、前述のようにドキュメントを選択し、それにスコアを付け、ランクを付けるであろう。一旦このランク付けが行なわれると、サーバ930は残りの検索ドキュメントをランク順にネットワーク接続925を介してクライアントPC920に送信し、そこで表示させるであろう。サーバはこれらのドキュメントを前述のようにユーザーの指示に従ってグループごとに送信するか、それらをグループごとに選択してクライアントPCで表示させるために全てのドキュメントを順次送信することができる。

【0062】

さらに、リモートコンピュータ（サーバ）930は、前述の従来の検索処理、自然言語処理および関連の処理の全てを与える1台のみのコンピュータによって実現される必要はなく、図9Dに示す分散処理方式であってもよい。その場合、このサーバが請け負う処理は分散処理方式における個別のサーバ間に分散される。ここで、サーバ930は、メッセージを接続950を介して（サーバ1、サーバ2、…、サーバnを含む）一連のサーバ960に分散するフロントエンドプロセッサ940からなる。これらのサーバの各々が本発明のプロセスの特定の部分

を実施する。この点で、サーバ1は後の検索のために入力ドキュメントの索引付けを行ない大容量データ記憶装置上のデータセットに格納するために用いることができる。サーバ2は、フロントエンドプロセッサ940によって送られるユーザから与えられるクエリに応答して大容量データ記憶装置から一組のドキュメントレコードを引出すための従来の統計的エンジンのようなサーチエンジンを実現できる。これらのレコードは、サーバ2からフロントエンドプロセッサ940を介してたとえばサーバnに送られ、対応のウェブサイトまたはデータベースからの対応の各ドキュメントを出力ドキュメント集合中にダウンロードするというような後処理が行なわれるであろう。フロントエンドプロセッサ940はまたクエリをサーバnに送るであろう。サーバnは次にそのクエリおよび各ドキュメントを適切に分析して論理形式三つ組の対応の集合を生じ、次に三つ組の集合を適切に比較し、前述のようにドキュメントを選択し、それにスコアをつけ、ランクをつけ、その後ランク付けされたドキュメントをフロントエンドプロセッサ940を介してクライアントPC920に戻して、そこでランク付けの表示がされるようにする。もちろん、本発明の処理において用いられるさまざまな動作は、実行時に生ずる条件および／または他により生ずる条件次第で、静的であれ動的であれ、他の多くの方法のうちの任意の方法によってサーバ960中に分散されてもよい。さらに、サーバ930は、たとえば従来のサーチエンジンのためのデータベースと自然言語処理のために用いられる辞書との両方が記憶され、サーバ内の全てのプロセッサからアクセス可能な共用直接アクセス記憶装置(DASD)である、たとえば周知のシスプレックス構成(または他の同様の分散マルチプロセス環境)によって実現することもできる。

【0063】

本発明を、検索された各ドキュメントレコードに응答してドキュメントをダウンロードし、次にそのレコードをたとえばクライアントPCによってローカルに分析してその対応の論理形式三つ組を生じるものとして説明したが、これに替えてこれらの三つ組はドキュメントに対しサーチエンジンが索引付けをしている間に生成されてもよい。この点で、サーチエンジンがたとえばウェブクローラを用いて、索引付けを行なうための新しい各ドキュメントを見つけたときに、エンジ

(54)

ンがそのドキュメントのための完全なファイルをダウンロードし、それからその直後またはさらに後にバッチ処理でそのドキュメントを分析し、その論理形式三つ組を生成することによってそのドキュメントを前処理することができる。前処理の終了時にサーチエンジンは次にこれらの三つ組をそのドキュメントのための索引付けされたレコードの一部としてそのデータベースに記憶するであろう。後に、そのドキュメントレコードがたとえばサーチクエリに応答して検索されるたびに、そのための三つ組がドキュメントレコードの一部として比較などの目的のためにクライアントPCに戻される。サーチエンジン内でのドキュメントの前処理によって、クライアントPCでのかなりの量の処理時間が節約されるという効果があり、それによってクライアントのスループットを増大させることができる。

【0064】

さらに、本発明をインターネットベースのサーチエンジンでの具体的な使用を例として説明したが、本発明は、(a) インターネットベースであろうとなかろうと、専用のネットワーク設備等によってアクセス可能な任意のネットワークアクセス可能なサーチエンジン、(b) それ自身が所持する予め記録されたデータセットに対して動作するローカルなサーチエンジン、たとえば、百科事典、年鑑または他の独立型スタンドアローンデータセットに代表されるCD-ROMベースのデータ検索アプリケーション、および／または(c) その任意の組合せでの使用に等しく適用可能である。

【0065】

上記を念頭において、図10Aおよび図10Bは、本発明のさらに他の実施例を集合的に示し、この実施例ではドキュメントの前処理によって論理形式三つ組を発生し、結果として生じる三つ組、ドキュメントレコードおよびドキュメント自体を独立型スタンドアローンデータセットとして既存の記憶媒体、たとえば、1つ以上のCD-ROMまたは(着脱可能なハードディスク、テープ、もしくは、光磁気または大容量磁気または電子記憶装置に代表される)他の可搬の大容量媒体にまとめて保存することによりエンドユーザへの頒布が容易になる。これらの図面の正しい配置は図10に示すとおりである。検索アプリケーション自体と

それに付随するサーチされるべきデータセットとを共通の媒体にまとめて入れることによって、スタンドアローンのデータ検索アプリケーションが得られ、それによって、ドキュメントを検索するためにリモートサーバにネットワーク接続することが必要でなくなる。

【0066】

図示したように、この実施例はドキュメント索引付け部分1005₁、複製部分1005₂およびユーザ部分1005₃の本質的に3つの部分からなる。部分1005₁はドキュメントを集め索引付けしてデータセット、すなわち図示するデータセット1030を作成し、データセット1030は、独立型ドキュメント検索アプリケーション、たとえば、百科事典、年鑑、(判例集のような)専用ライブラリ、定期刊行物のコレクション等のためのドキュメントリポジトリを形成する。大記憶容量を有するCD-ROMおよび他の形態の媒体を複製するためのコストは急速に低下しつつあり、この実施例は大量のドキュメントをそれを正確にサーチする性能とともに広いユーザコミュニティに費用効率よく頒布するために特に魅力的である。

【0067】

いずれにせよ、索引付けされデータセットを形成するために入力されるドキュメントは任意数の多様なソースから集められ、コンピュータ1010に順次与えられる。このコンピュータはメモリ1015内に記憶されている適切なソフトウェアによってドキュメント索引付けエンジンを実現し、ドキュメント索引付けエンジンは、このような各ドキュメントのためのレコードをデータセット1030内に作成し、そのドキュメントのためのレコードに情報を保存し、またドキュメント自体のコピーを含む適切なエントリをデータセット内に作成し保存する。エンジン1015は三つ組発生プロセス1100を実行する。図11に関連して以下に詳細に説明するこのプロセスは索引付けされる各ドキュメントごとに別個に実行される。本質的には、このプロセスは、図6Aおよび図6Bに示すブロック640に対して前述したのと本質的に同じように、ドキュメント内のテキスト句を分析し、そうすることによってそのドキュメントに対する論理形式三つ組の対応の集合を構成してデータセット1030内に記憶する。図10Aおよび図10

Bに示す索引付けエンジン1010によって実行されてドキュメントに索引を付ける他の全てのプロセス、たとえばそのための適切なレコードを発生するプロセスはいずれも本発明には無関係であるので、それらについては詳細に述べない。三つ組の集合がプロセス1100によって一旦発生されると、エンジン1015がこの集合をドキュメント自体のコピーとそれに対して作られたドキュメントレコードとともにデータセット1030へと記憶する、と述べるだけで十分である。したがって、全索引付け動作が終わると、データセット1030は索引付けされた全ドキュメントの完全なコピーとそのためのとをその中に記憶しているだけでなく、そのドキュメントのための論理形式三つ組の集合をも記憶している。

【0068】

一旦所望の全ドキュメントが適切に索引付けされると、データセット1030は、これは「マスタデータセット」と見ることができるが、次に複製部分1005₂によって複製される。部分1005₂内では、従来の媒体複製システム1040が線1035によって供給されるマスタデータセットの内容のコピーを、線1043によって供給される検索プロセスおよびユーザインストールプログラムを含む検索ソフトウェアのための適切なファイルのコピーとともに、1つ以上のCD-ROMのような共通の記憶媒体に繰返し書込んで、スタンドアローンのドキュメント検索アプリケーションをまとめて形成する。システム1040によって、個々の複製1050₁、1050₂、…、1050_nを有する一連の媒体複製1050が生成される。全複製は同一であり、複製1050₁に関して具体的に示してあるように、線1043によって供給されるドキュメント検索アプリケーションファイルのコピーと線1035によって供給されるデータセット1030のコピーとを含む。データセットのサイズおよび構成次第では、各複製が1つ以上の別個の媒体、たとえば別個のCD-ROMにまたがってもよい。後に、複製は典型的にはライセンスの取得によって破線1055で示すようにユーザコミュニティ中に流通される。

【0069】

一旦ユーザ、たとえばユーザ_jがユーザ部分1005₃に示すように(CD-ROM1060とも示す)CD-ROM_jのような複製を入手すると、ユーザは、

(57)

(同一の構成でないとしても実質的に図 3 に示すクライアント PC 300 のような構成を有する PC のような) コンピュータシステム 1070 によって、本発明を含むドキュメント検索アプリケーションを CD-ROM_j に記憶されているデータセットに対して実行してそこから所望のドキュメントを引出すことができる。特に、ユーザは CD-ROM_j を入手した後、CD-ROM を PC 1070 に挿入し、CD-ROM に記憶されているインストールプログラムの実行を始め、それによって、ドキュメント検索アプリケーションファイルのコピーを作り、それを PC のメモリ 1075、通常はハードディスク内の予め規定されたディレクトリへとインストールし、それによって、PC 上にドキュメント検索アプリケーション 1085 を作成する。このアプリケーションはサーチエンジン 1090 および検索プロセス 1200 を含む。一旦インストールが完了し、アプリケーション 1085 が呼出されると、ユーザは適切なフルテキストのクエリをアプリケーションに与えることによって、CD-ROM_j のデータセットをサーチすることができる。クエリに応答して、サーチエンジンはそれらのドキュメントのためとこのような各ドキュメントのための記憶されている論理形式三つ組とを含むドキュメントの集合をデータセットから引出す。クエリは検索プロセス 1200 にも与えられる。このプロセスは図 6 A および図 6 B に関連して前述した検索プロセス 600 に非常に類似しており、クエリを分析し、そのため論理形式三つ組を構成するものである。その後、図 10 A および図 10 B に示すプロセス 1200 がその集合内の検索されたドキュメントの各々のための論理形式三つ組、特にそのためのレコードをクエリのための三つ組と比較する。それらの間で発生する三つ組の一致とそれらの重みとに基づき、プロセス 1200 は詳細に前述した態様で少なくとも 1 つの一致する三つ組を示すドキュメントの各々をスコア付けし、これらのドキュメントを降順のスコアでランク付けし、最後に、最も高いランク付けを有する典型的に 5 - 20 またはそれよりも少ない小グループのドキュメントレコードをユーザに視覚的に提示する。ユーザはこれらのを検討し、関連のあるように思われる任意のドキュメントのコピー全体を検索し表示するようドキュメント検索アプリケーションに指示することができる。一旦ユーザが最初のグループの検索ドキュメントに対する最初のグループのドキュメントレコードを検討

(58)

すると、ユーザは次に高いランク付けを有する次のグループのドキュメントレコードを要求することができ、以下同様に、検索された全ドキュメントレコードを検討し終わるまでこれを行なうことができる。アプリケーション1085は、初期状態では、ランク付けされたドキュメントレコードをクエリに応答して戻すが、これに替えてこのアプリケーションがドキュメント自体のランク付けされたコピーをクエリに応答して戻してもよい。

【0070】

図11は、図10Aおよび図10Bに示すドキュメント索引付けエンジン1015によって行なわれる三つ組発生プロセス1100を示す。前述のように、プロセス1100は索引付けされるべきドキュメントの前処理を、そのドキュメントにおけるテキストフレーズを分析し、そうすることによってそのドキュメントのための論理形式三つ組の対応の集合を構成してデータセット1030内に記憶することによって行なう。特に、プロセス1100を開始するとブロック1110が実行される。このブロックは初めに、そのドキュメントに関連したHTMLタグ内にある任意のテキストを含む全テキストをそのドキュメントから抽出する。その後、一度に1文ずつ行なわれる自然言語処理を容易にするために、各ドキュメントのためのテキストが従来の一文切出し処理によって分解され、各文（または疑問文）がファイル内で別個のラインを占めるテキストファイルとなる。その後、ブロック1110が（図13Aに関連して詳細に後述する）NLPルーチン1300をそのドキュメント内のテキストの各ラインごとに別個に呼出して、このドキュメントを分析し、そのラインのための論理形式三つ組の対応の集合を構成してデータセット1030内にローカルに記憶する。これらの動作が完了すれば、ブロック1110およびプロセス1100の実行が終了する。

【0071】

図10Aおよび図10Bに示す本発明の具体的な実施例において用いられるような本発明の検索プロセス1200のフローチャートを図12Aおよび図12Bに集合的に示す。図12Aおよび図12Bの図面の正しい配列は図12に示すとおりである。（図6Aおよび図6Bに示し、詳細に前述した）検索プロセス600とは対照的に、図12Aおよび図12Bに示す全動作は共通のコンピュータシ

システム、ここではPC1070（図10Aおよび図10B参照）において行なわれる。理解を助けるため、以下の説明においては図10Aおよび図10Bを同時に参照されたい。

【0072】

プロセス1200を開始すると、ブロック1205が初めに実行される。このブロックは実行されるとユーザにフルテキストのクエリを入力させる。一旦このクエリが得られると、実行経路は分岐して経路1207によってブロック1210へ、および経路1243によって経路1245へ進む。ブロック1245は実行されるとNLPルーチン1350を呼出してクエリを分析し、対応の論理形式三つ組の集合を構成し、それをローカルにメモリ1075内に記憶する。ブロック1210は実行されると、破線1215で示すようにフルテキストのクエリをサーチエンジン1090に送信する。この時点で、サーチエンジンはクエリに回答してブロック1220を実行して、ドキュメントレコードの集合とこのようなレコードの各々に関連した関連の論理形式三つ組との両方を検索する。この集合と関連の論理形式三つ組とが検索されれば、両方は破線1230で示すようにプロセス1200に与えられ、具体的にはそこにおけるブロック1240に与えられる。ブロック1240は単にこの情報をサーチエンジン1090から受け、それを後に使用するためにメモリ1075内に記憶する。ブロック1245における動作をブロック1210、1090および1220における動作と本質的に並列的に行われるものと説明したが、ブロック1245における動作は実際の実行上の観点からブロック1210、1090または1220における動作の前または後に直列的に行なわれてもよい。

【0073】

論理形式三つ組の集合がクエリと検索された各ドキュメントレコードとの両方のためにメモリ1075へと記憶されれば、ブロック1250が実行される。このブロックは詳細に前述した態様で、クエリ内の論理形式三つ組の各々を、検索された各ドキュメントレコードのための論理形式三つ組の各々と比較して、クエリ内の任意の三つ組と対応の任意のドキュメントの任意の三つ組との間の一致を突き止める。一旦ブロック1250が完了すると、ブロック1255が実行され

(60)

て、一致する三つ組を示さない、すなわち、クエリ内の任意の三つ組と一致する三つ組を有さないドキュメントに対する検索された全レコードを廃棄する。その後、ブロック 1 2 6 0 が実行される。ブロック 1 2 6 0 によって、残る全ドキュメントレコードが、前述のように、対応の各ドキュメントごとに存在する一致する三つ組の関係のタイプとそれらの重みとに基づいてスコアを割当てられ、ドキュメントレコードがそのように重み付けされれば、ブロック 1 2 6 5 が実行されてスコアの降順にレコードをランク付ける。最後に、ブロック 1 2 7 0 が実行されて、典型的には最も高いスコアを示す予め規定された小グループ、典型的には 5 または 1 0 のドキュメントレコードについてレコードをランク順に表示する。その後、ユーザはたとえばコンピュータシステム 1 0 7 0 によって表示されている対応のボタンの上でマウスを適切に「クリックする」ことによって、ランク付けされたドキュメントレコードの次のグループをそのシステムに表示させ、以下同様にユーザがランク付けされた全ドキュメントレコードを順に十分に調べる（そしてその中の関心のある任意のドキュメントにアクセスし、それを調べる）までその動作を行なう。この時点で、プロセス 1 2 0 0 は完了され、実行が終了する。

【0 0 7 4】

図 1 3 A は、図 1 1 に示す三つ組発生プロセス 1 1 0 0 内で実行される NLP ルーチン 1 3 0 0 のフローチャートを示す。前述のように、NLP ルーチン 1 3 0 0 は索引付けされるべき入来するドキュメント、具体的にはそのためにテキストの 1 ラインを分析し、そのドキュメントのための論理形式三つ組の対応の集合を構成し、それをローカルに図 1 0 A および図 1 0 B に示すデータセット内に記憶する。ルーチン 1 3 0 0 は、図 7 に示し、詳細に前述した NLP ルーチン 7 0 0 と本質的に同様に動作する。

【0 0 7 5】

特に、ルーチン 1 3 0 0 が開始されると、ブロック 1 3 1 0 が最初に実行されて、入力テキストのラインを処理して図 5 A に示す例示のグラフ 5 1 5 のような論理形式グラフを生成する。その後、図 1 3 A に示すように、ブロック 1 3 2 0 が実行されてそのグラフから対応の論理形式三つ組の集合を抽出する（読出す）

(61)

。一旦これが起こると、ブロック1330が実行されて別個に、かつ区別してフォーマット化されたテキスト文字列としてこのような論理形式三つ組の各々を生成する。最後に、ブロック1340が実行されて、入力されたテキストのそのラインと、一連のフォーマット化されたテキスト文字列として、そのラインのための論理形式三つ組の集合とがデータセット1030に保存される。この集合が完全に記憶されれば、ブロック1300の実行を終了する。これに替えて、論理形式三つ組ではなく異なる形式、たとえば論理形式グラフまたはそのサブグラフが本発明に関連して用いられるのであれば、ブロック1320および1330を、フォーマット化された文字列としてその特定の形式を発生するように、そしてブロック1340が論理形式三つ組の代わりにその形式をデータセットに記憶するように、容易に変更できるであろう。

【0076】

図13Bは、検索プロセス1200内で実行されるNLPルーチン1350のフローチャートを示す。前述のように、NLPルーチン1350はユーザ_jによって（図10Aおよび図10Bに示す）ドキュメント検索アプリケーション1085に与えられるクエリを分析し、そのための対応の論理形式三つ組の集合を構成し、メモリ1075内にローカルに記憶する。図13Aに関連して詳細に前述したルーチン1300とルーチン1350との間の動作上の唯一の違いは、対応の三つ組が記憶される場所である。すなわち、NLPルーチン1300におけるブロック1340の実行ではデータセット1030に記憶され、NLPルーチン1350におけるブロック1390の実行ではメモリ1075に記憶される、という点である。ルーチン1350の他のブロック、具体的にはブロック1360、1370および1380によって行なわれる動作はルーチン1300のブロック1310、1320および1330によってそれぞれ行われるのと実質的に同じであるので、前者のブロックの詳細な説明を省略する。

【0077】

図1Aに関連して一般的に前述したような本発明の検索プロセスの性能を試験的に試すために、本発明の検索システムにおいてサーチエンジンとしてALTA VISTAサーチエンジンを用いた。インターネット上で誰もがアクセス可能なこのエン

(62)

ジンは3100万を超えるウェブページが索引付けされていると称されている従来の統計的サーチエンジンであり、広く用いられている（概算で現在1日当たり2800万ヒットを記録している。）。本発明の検索プロセス600を、MICROSOFT OFFICE 97プログラムスイートの一部を成す文法チェッカー内に含まれる、辞書ファイルを含むさまざまな自然言語処理コンポーネントを用いて、一般的なPentium 90 MHzのPC上で実現した（「OFFICE」および「OFFICE 97」はワシントン州レッドモンドのMicrosoft Corporationの商標である）。我々はオンラインのパイプライン処理モデルを用いた。すなわち、続く結果をユーザが待っている間にドキュメントがパイプラインの態様でオンラインで集められ、処理された。この特定のPCは、各センテンスごとに論理形式三つ組を発生するのに約3分の1秒から約2分の1秒を要した。

【0078】

サーチエンジンに与えるためのフルテキストのクエリを作るようボランティアに依頼した。合計121個の広範囲の互いに異なるクエリが作られ、その代表的なものは「Why was the Celtic civilization so easily conquered by the Romans? (なぜケルト文明はローマ人によって簡単に征服されたのか?)」、「Why do antibiotics work on colds but not on viruses? (抗生物質はなぜ風邪に効くのにウィルスには効かないのか?)」、「Who is the governor of Washington? (ワシントン州の知事は誰か?)」、「Where does the Nile cross the equator? (ナイル川が赤道と交差するのはどこか?)」、および「When did they start vaccinating for small pox? (天然痘の予防接種が始められたのはいつ頃か?)」といったものであった。これらの121個の各クエリをALTA VISTAサーチエンジンに与え、各クエリに応答して戻った利用可能なものからなる上位30のドキュメントを得た。クエリの中には30未満のドキュメントしか戻らないものがあり、その場合は戻った全ドキュメントを使用した。全121クエリに対して、延べ3361ドキュメント（すなわち、「生の」ドキュメント）が得られた。

【0079】

3361のドキュメントと121のクエリとの各々が本発明のプロセスによって分析されて論理形式三つ組の対応の集合が生成された。集合は適切に比較され

(63)

、結果のドキュメントが前述のように選択され、スコアおよびランクをつけられた。

【0080】

3361のドキュメントの全てを、そのドキュメントが得られた対応のクエリとの関連性について手作業で別個に評価した。関連性を評価するため、発明者の具体的な実験目的を知らない一人の人を評価者として利用し、これらの3361ドキュメントの各々をその対応のクエリとの関連について「最適」、「関連あり」または「関連なし」として手作業で主観的にランク付けした。最適なドキュメントとは、対応のクエリに対する明らかな回答を含むものであるとされた。関連のあるドキュメントとは、クエリに対する明らかな回答を含まないがそれにもかかわらず関連性のあるもののことであるとされた。関連のないドキュメントとは、クエリに対する有益な回答ではないもの、すなわち、クエリに関連がないか、英語以外の言語によるか、またはALTA VISTAエンジン（すなわち、「cobweb」リンク）によって与えられた対応のURLからは検索できないドキュメントのことであるとされた。評価精度を高めるため、二人目の評価者がこれらの3361ドキュメント内のサブセット、具体的には、その対応のクエリにおける論理形式三つ組と一致する少なくとも1つの論理形式三つ組を示したドキュメント（3361ドキュメントのうち431）と、それまでに関連ありまたは最適であるとランク付けされたが一致する論理形式三つ組を有さないドキュメント（3361ドキュメントのうち102）とを調査した。ドキュメントに対するこれらのランク付けの意見の相違があれば、それは「仲裁者」となる三人目の評価者によって検討された。

【0081】

この実験の結果として、関連した全ドキュメントにおいて、本発明の検索システムは、ALTA VISTAサーチエンジンが戻す生のドキュメントよりも改善を示したことが観察された。全体の（すなわち、選択された全ドキュメントの）精度では約16%から約47%へと200%程度改善し、上位5ドキュメント内では約26%から約51%へと約100%改善した。加えて、本発明のシステムの使用によって、最適であるとして戻された最初のドキュメントの精度は生のドキュメン

(64)

トのそれに対して約 17% から約 35% へと約 113% 改善された。

【0082】

本発明を統計的サーチエンジンでの使用を例として具体的に説明したが、本発明はそれに限定されない。その点では、本発明は実質的にいかなるタイプのサーチエンジンによって得られた検索ドキュメントをも処理してそのエンジンの精度を高めるよう利用することができる。

【0083】

論理形式三つ組内の各種の属性ごとに固定した重みを用いる代わりに、これらの重みを動的に変化させ、実際のところ、適応的としてもよい。これを達成するため、たとえばベイズのネットワークまたはニューラルネットワークのような学習メカニズムを本発明のプロセスに適切に組み込み、各種の論理形式三つ組のための重み数値を学習経験に基づいて最適な値に変えてもよい。

【0084】

本発明のプロセスは論理形式三つ組が正確に一致することを必要としたが、三つ組の間で十分に類似した意味内容を識別する目的のために、一致判断の基準を緩和してパラフレーズも一致とみなすようにしてもよい。パラフレーズは語彙的であってもよく構造的であってもよい。語彙的パラフレーズの例は上位語または類義語であろう。構造的パラフレーズの例は同格関係にある名詞または関係節の使用である。たとえば、「大統領、ビル・クリントン」のような同格関係にある名詞構成は「大統領であるビル・クリントン」のような関係節構成と一致するものとみなされるべきである。意味レベルでは、2つの語が互いにいかに意味的に類似しているかについてきめの細かい判断をすることができ、それによって、「どこでコーヒーが栽培されるか」というクエリと「コーヒーは熱帯山岳地方でよく栽培される」というようなコーパス（文例）文との間の一致を認めることができる。加えて、一致が存在するか否かを判断するための手順を、与えられたクエリのタイプによって変更することができる。たとえば、何かが存在する場所についてクエリが尋ねていれば、その手順は、ある文がクエリと一致するとみなされるためには、テストされている文と関連した任意の三つ組内に「場所」属性が存在していなければならない、と要求するであろう。したがって、論理形式三つ組の

「一致」とは同一の一致だけをいうのみならず、このような緩和された、判断を含むような、変更された一致条件の全てから生じるものをも包含するように包括的に規定される。

【0085】

さらに、本発明をグラフィックス、表、映像またはその他のような非テキスト情報の検索を中心とする他の処理技術と容易に組合せて全体の精度を高めることができる。一般に、ドキュメント中の非テキスト内容にはよく、たとえば図の記号または短い説明のような、そのドキュメント中の言語的（テキストによる）描写が付随するものである。したがって、本発明のプロセス、特にその自然言語成分の使用を、非テキスト内容にしばしば付随する言語的描写を分析し、処理するために用いることができる。クエリに意味的に関連した言語的内容を示すドキュメントの集合を初めに探し、次に、このドキュメントの集合をそれらの非テキスト内容に関して処理することによって、本発明の自然言語処理技術を用いて関連あるテキスト内容および非テキスト内容を有するドキュメントを検索することができる。これに替えて、ドキュメント検索を初めに非テキスト内容について行なってドキュメントの集合を取出し、次に本発明の技術によってそのドキュメントの集合をそれらの言語的内容について処理することで関連のあるドキュメントを検索してもよい。

【0086】

本発明の教示を採用したさまざまな実施例を図示し、詳細に説明したが、当業者であればこれらの教示をなお利用する多くの他の実施例に容易に想到することができるであろう。

【図面の簡単な説明】

【図1】 我々の本発明に従う情報検索システム5の非常に高いレベルのブロック図である。

【図2】 我々の本発明の教示を利用する、図1に示されるタイプの情報検索システム200の高いレベルの実施例を示す図である。

【図3】 図2に示されるシステム200内に含まれる、特定的にはクライアントパーソナルコンピュータであるコンピュータシステム300を示すブロッ

ク図である。

【図4】 図3に示されるコンピュータ300内で実行されるアプリケーションプログラム400を示す非常に高いレベルのブロック図である。

【図5】 AからDは、種々の複雑さを有する英語の文の種々の対応例、およびそれらに関する対応の論理形式要素を示す図である。

【図6】 図6は図6Aおよび図6Bの図面の正しい配置を示す図であり、図6Aおよび図6Bは、我々の発明の検索プロセス600のフローチャートを合わせて示す図である。

【図7】 プロセス600内で実行されるNLPルーチン700のフローチャートを示す図である。

【図8A】 例として、一致する論理形式三つ組の重み付け表800を示す図である。

【図8B】 例示的な質問および統計的に検索された例の3つのドキュメントの組に関する、図6Aおよび図6Bに全て示されているブロック650、660、665および670内で行なわれる、我々の発明の教示に従う論理形式三つ組の比較、ドキュメントのスコア付け、ランク付けおよび選択処理を視覚的に示す図である。

【図9】 AからCは、我々の本発明の教示を採用した情報検索システムの3つの異なった実施例をそれぞれ示す図であり、Dは我々の本発明のさらに別の異なった実施例を実現するにあたり使用される、図9Cに示されるリモートコンピュータ（サーバ）930の代替的な実施例を示す図である。

【図10】 図10は図10Aおよび図10Bの図面の正しい配置を示す図であり、図10Aおよび図10Bは我々の本発明のさらに別の実施例であって、各ドキュメントに関する論理形式三つ組が、それらに関するドキュメントレコードとともに予め計算されて記憶され、後のドキュメント検索動作時にアクセスされるものを合わせて示す図である。

【図11】 図10Aおよび図10Bに示されるドキュメント索引付けエンジン1015によって行なわれる三つ組生成処理1100を示す図である。

【図12】 図12は図12Aおよび図12Bの図面の正しい配置を示す図

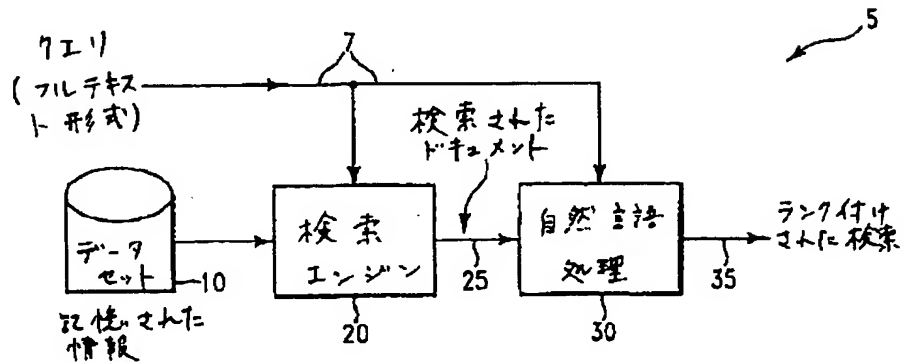
(67)

であり、図12Aおよび図12Bは、図10Aおよび図10Bに示されるコンピュータシステム300内で実行される我々の発明の検索処理1200のフローチャートを合わせて示す図である。

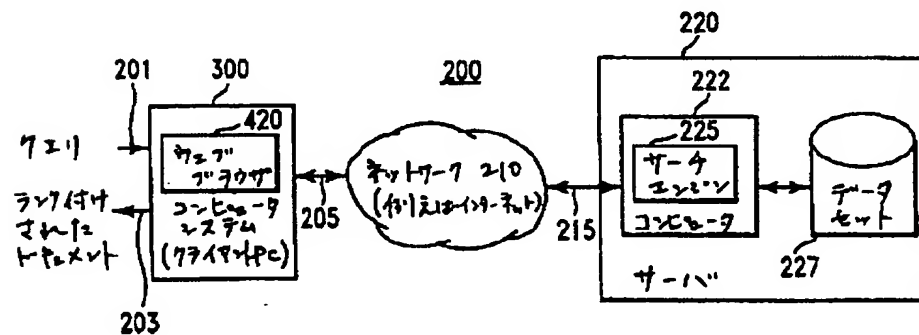
【図13A】 三つ組生成処理1100内で実行されるNLPルーチン1300のフローチャートを示す図である。

【図13B】 検索処理1200内で実行されるNLPルーチン1350のフローチャートを示す図である。

【図1】

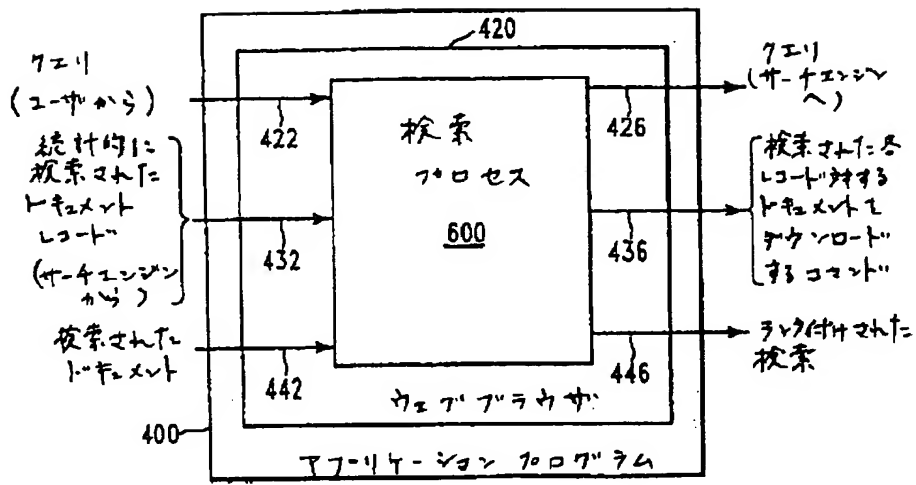


【図2】



(69)

【図 4】



【図 5 A】

510 入力文字列: THE OCTOPUS HAS THREE HEARTS.

論理形式:
グラフ

```

  HAVE
  ├── Dsub — OCTOPUS
  └── Dobj — HEART
              └── Ops — THREE
  
```

515

論理形式:
三つ組

```

  HAVE — Dsub — OCTOPUS
  HAVE — Dobj — HEART
  HEART — Ops — THREE
  
```

525

【図 5 B】

530 入力文字列: THE OCTOPUS HAS THREE HEARTS AND TWO LUNGS.

論理形式:
グラフ

```

  HAVE
  ├── Dsub — OCTOPUS
  └── Dobj — AND
                ├── Crds — HEART
                │       └── Ops — THREE
                └── LUNG
                        └── Ops — TWO
  
```

535

論理形式:
三つ組

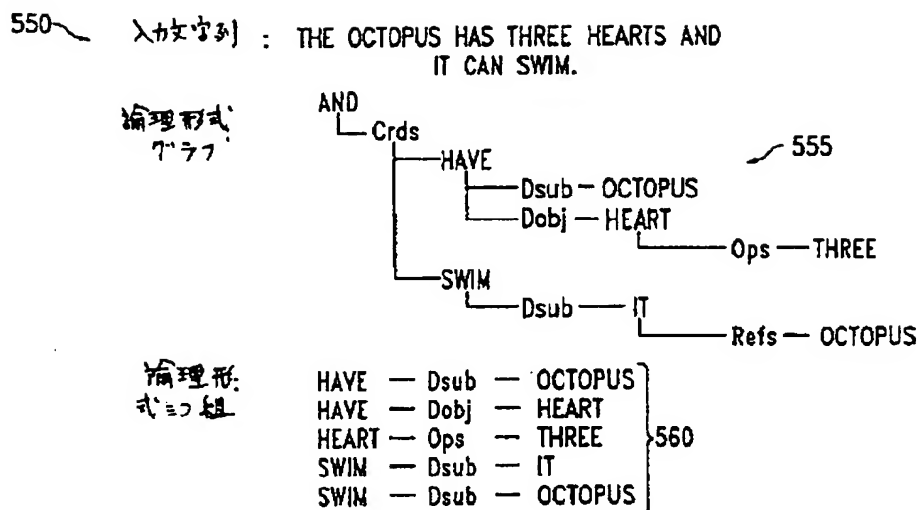
```

  HAVE — Dsub — OCTOPUS
  HAVE — Dobj — HEART
  HAVE — Dobj — LUNG
  HEART — Ops — THREE
  LUNG — Ops — TWO
  
```

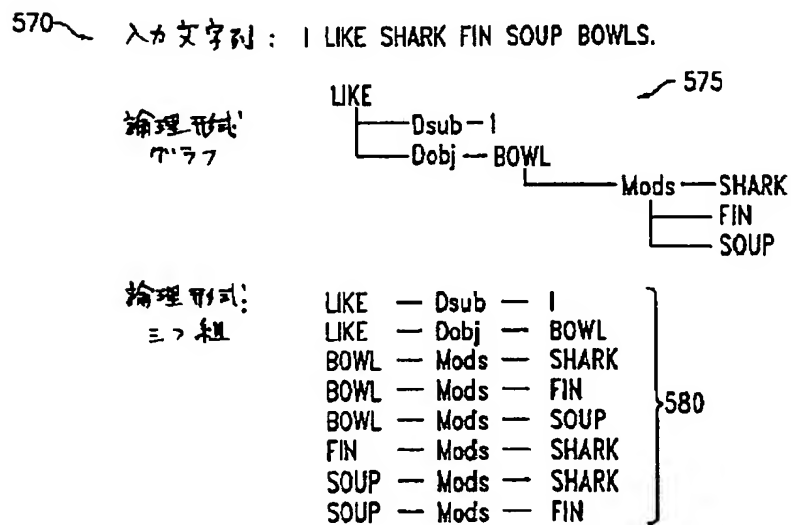
540

(70)

【図 5 C】



【図 5 D】



【図 6】

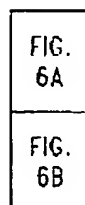
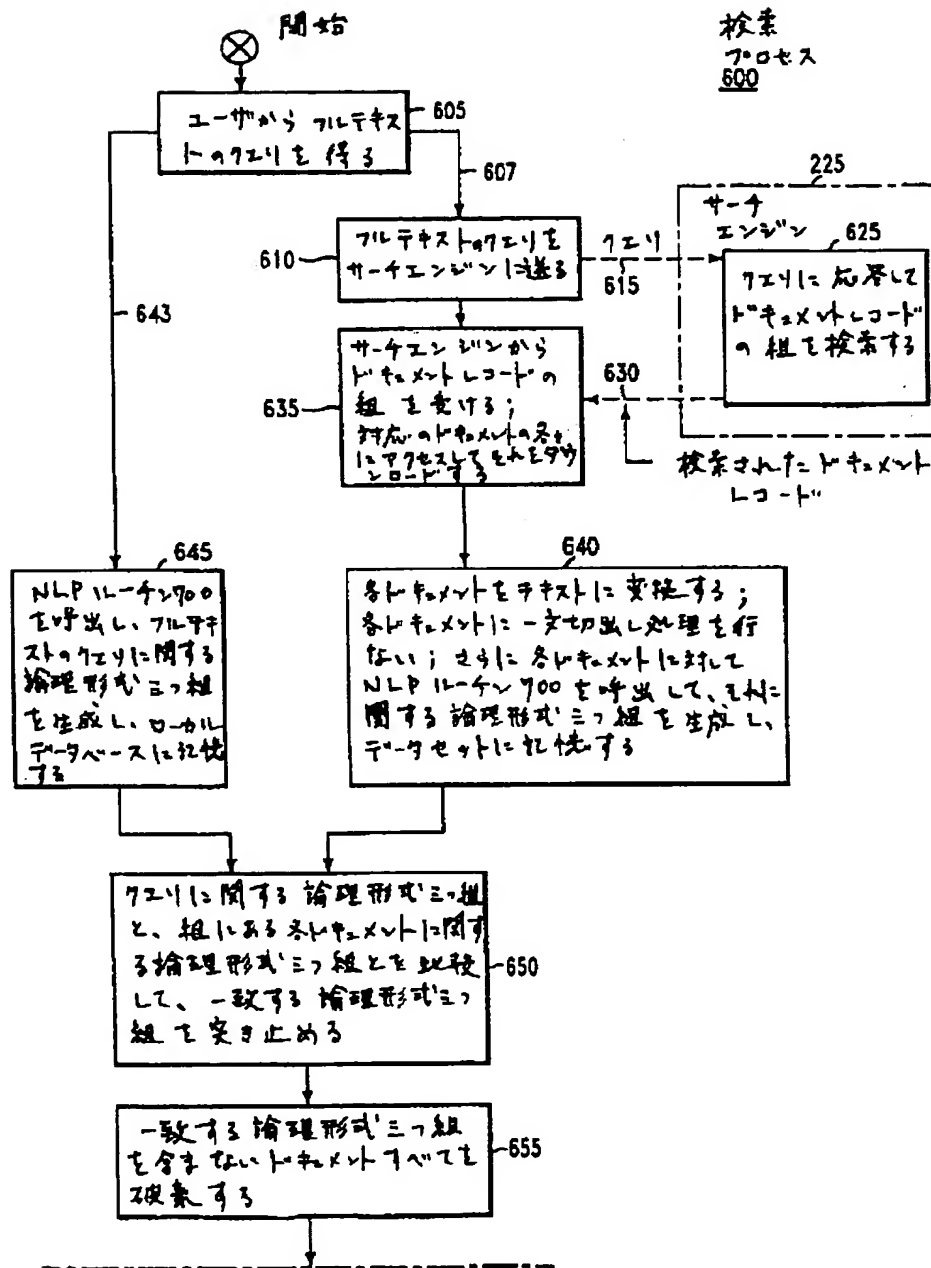


FIG. 6

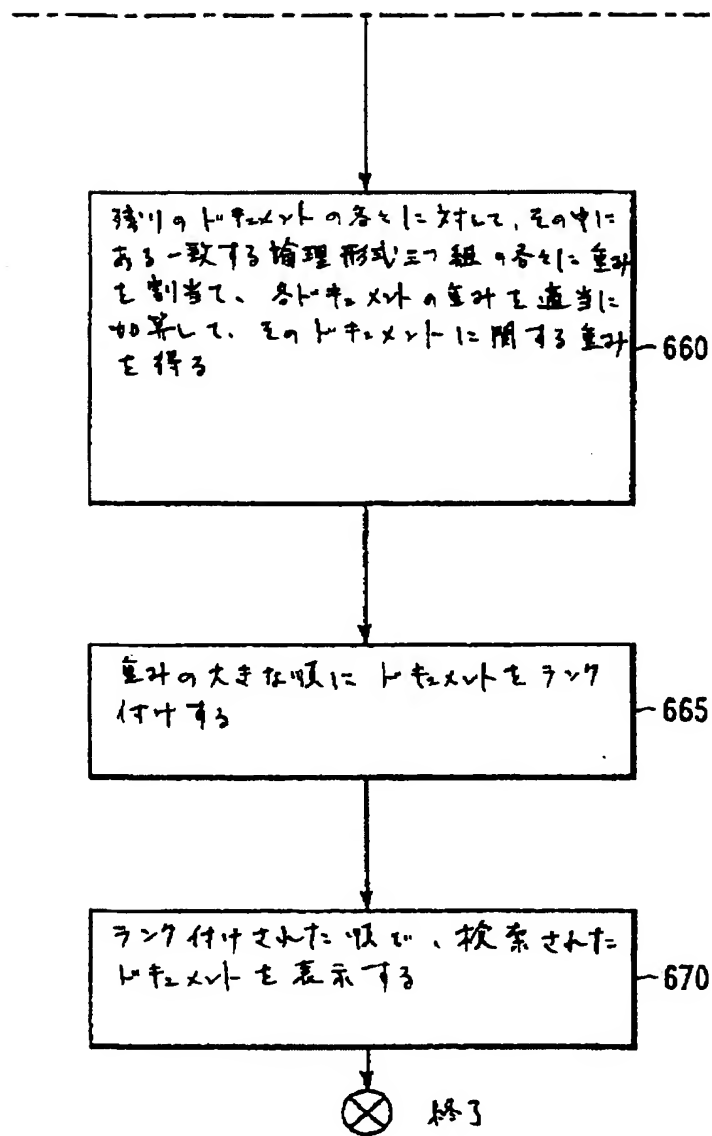
(71)

【図 6 A】



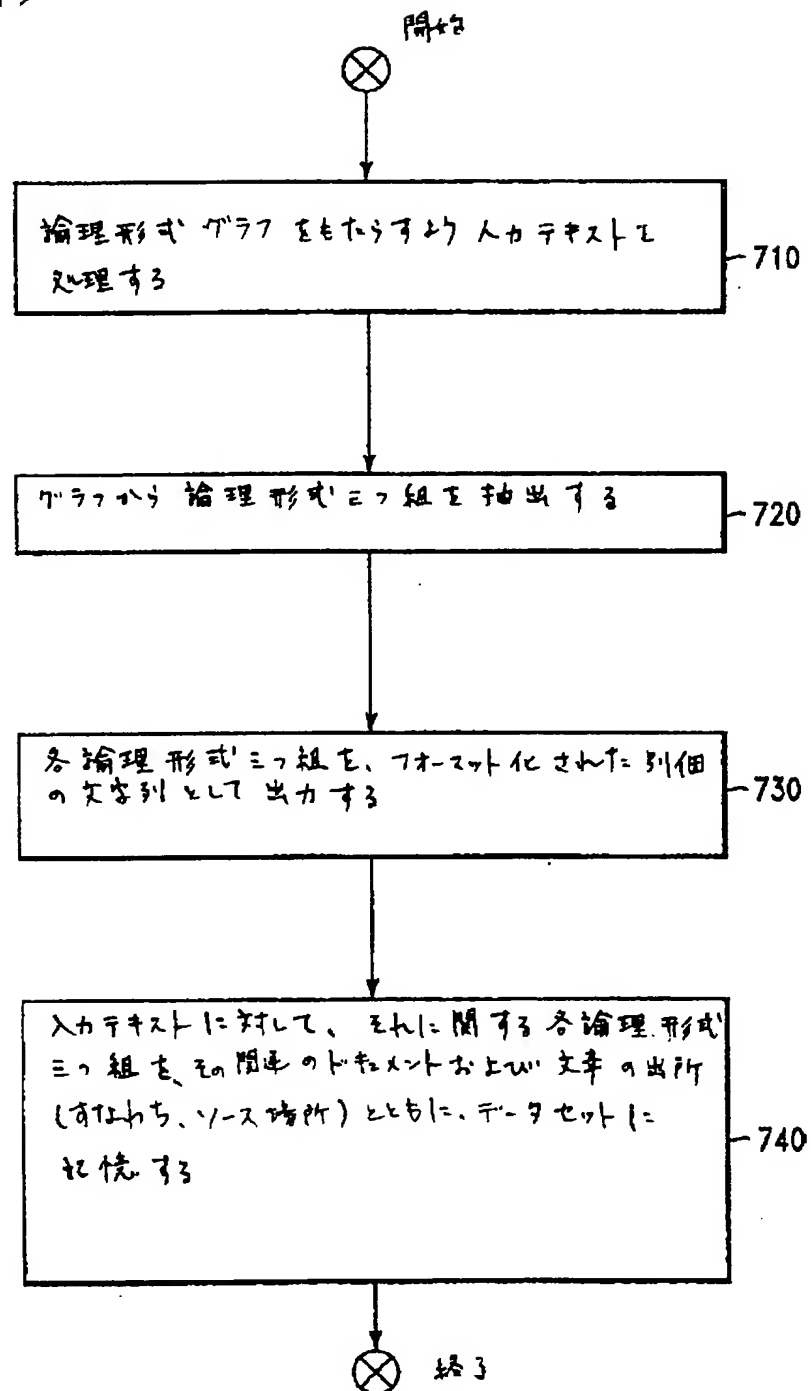
(72)

【図 6 B】



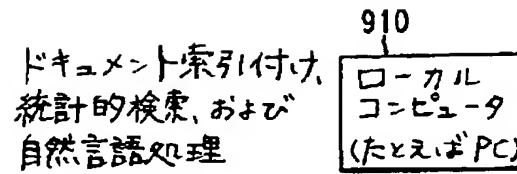
(73)

【図 7】

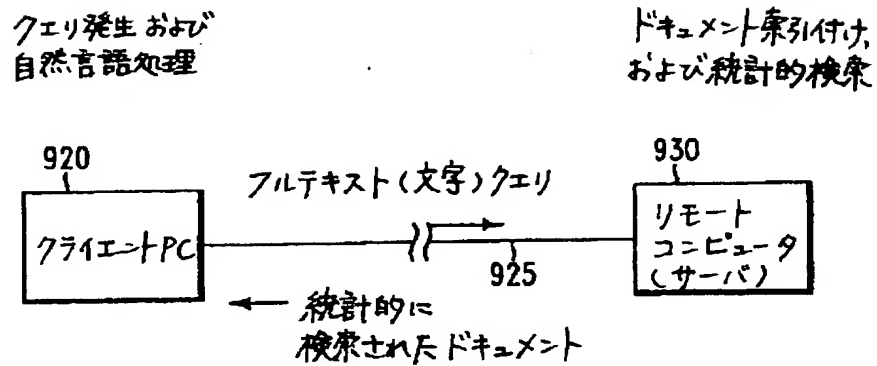
NLP 処理
700

(75)

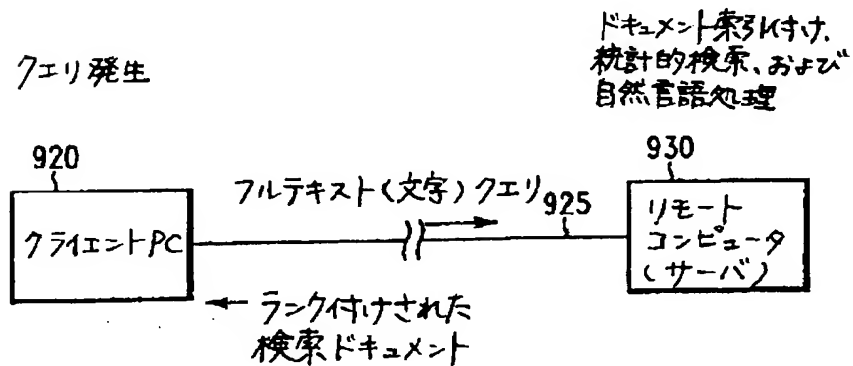
【図 9 A】



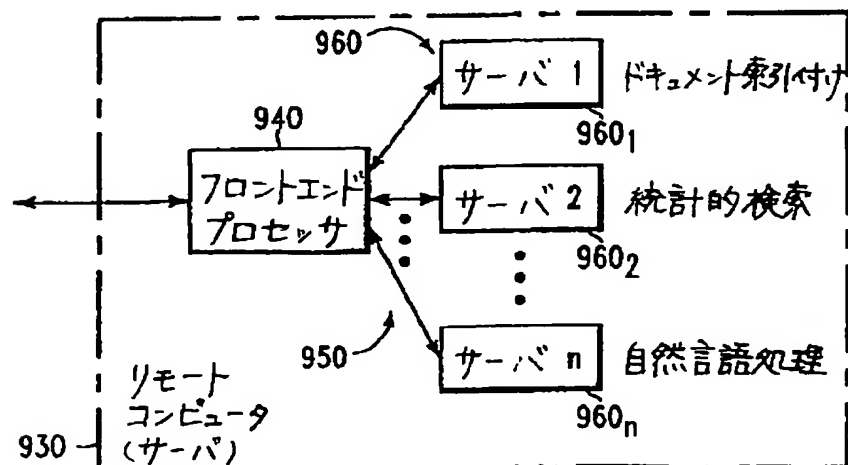
【図 9 B】



【図 9 C】



【図 9 D】



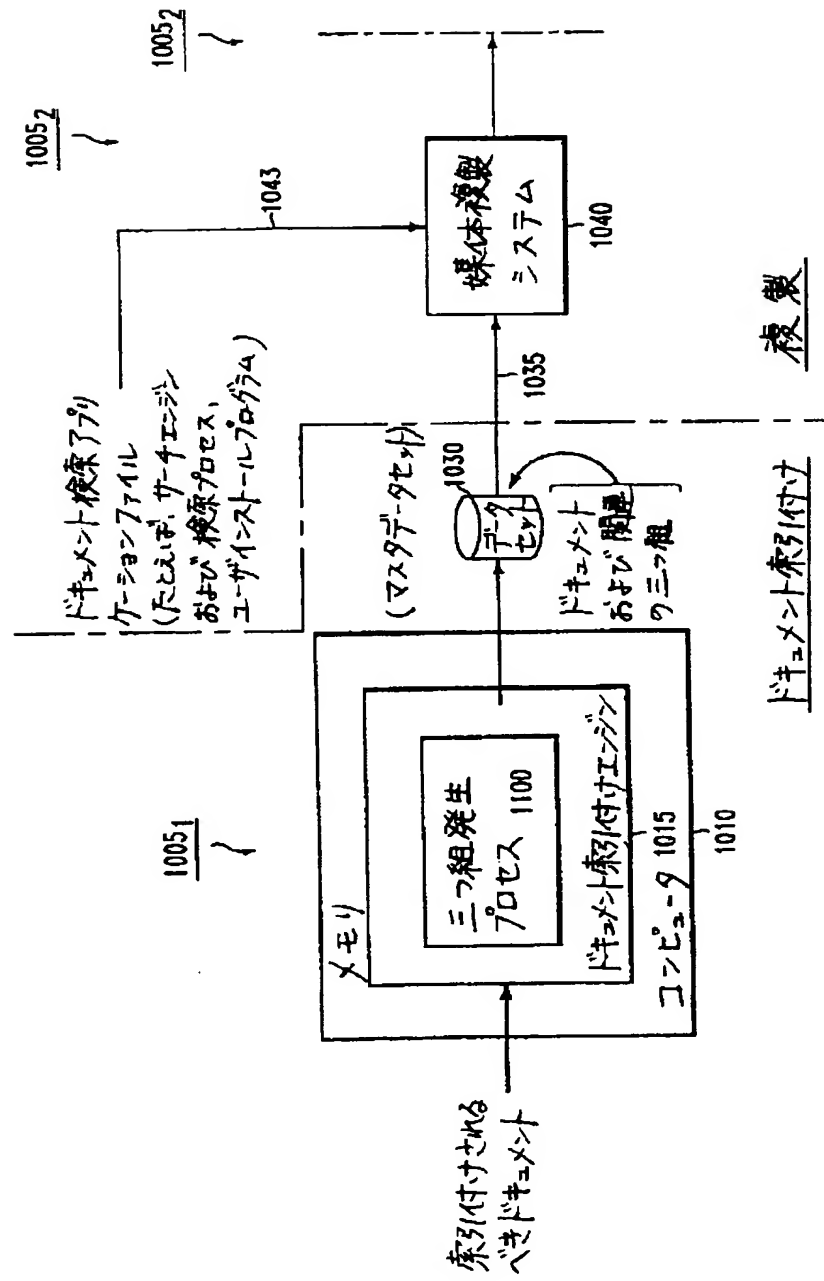
(76)

【図 10】



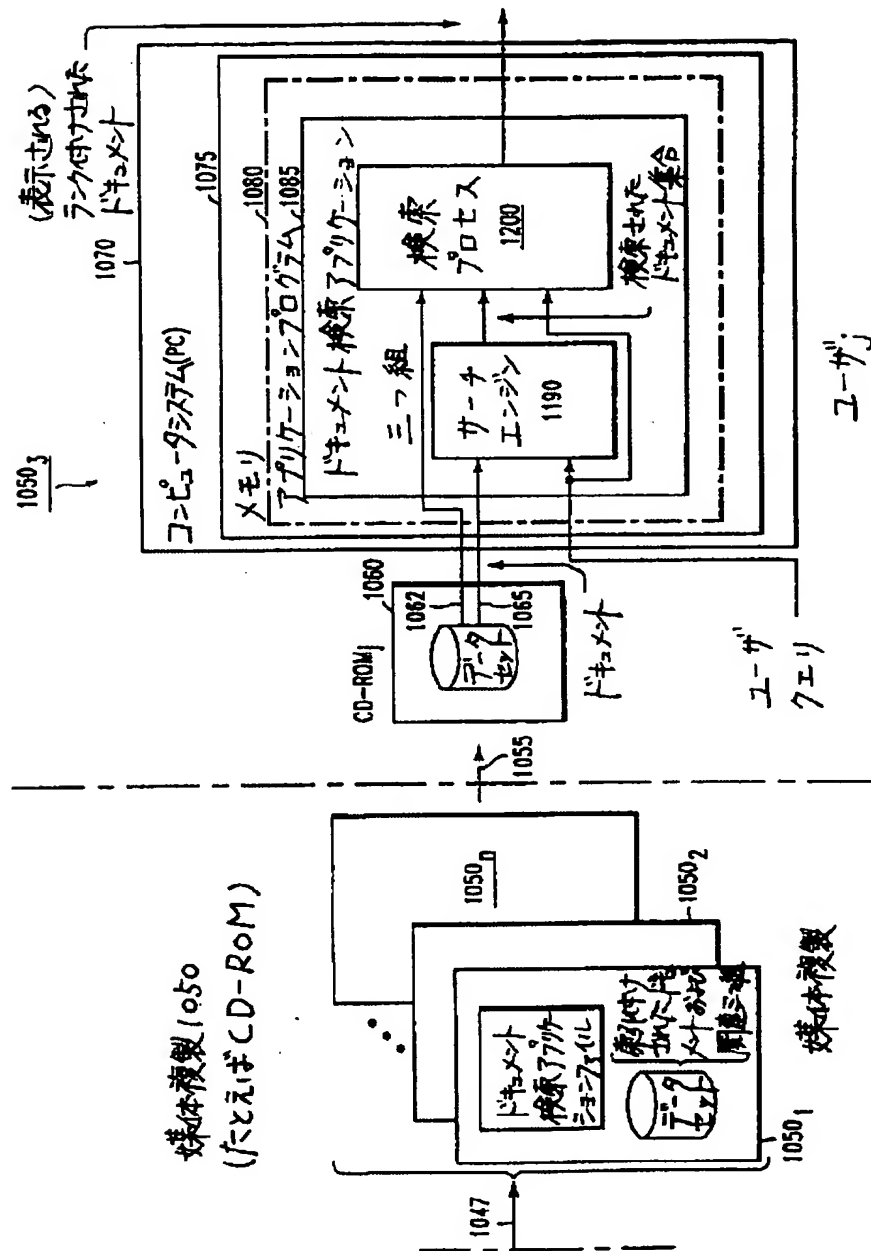
(77)

【図10A】



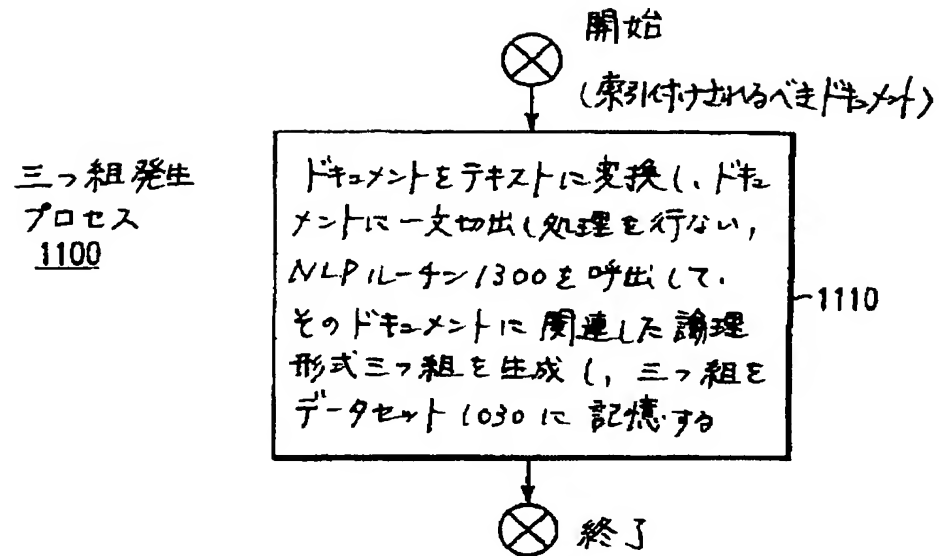
(78)

【図10B】



(79)

【図 11】



【図 12】

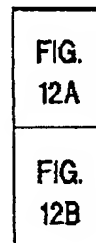
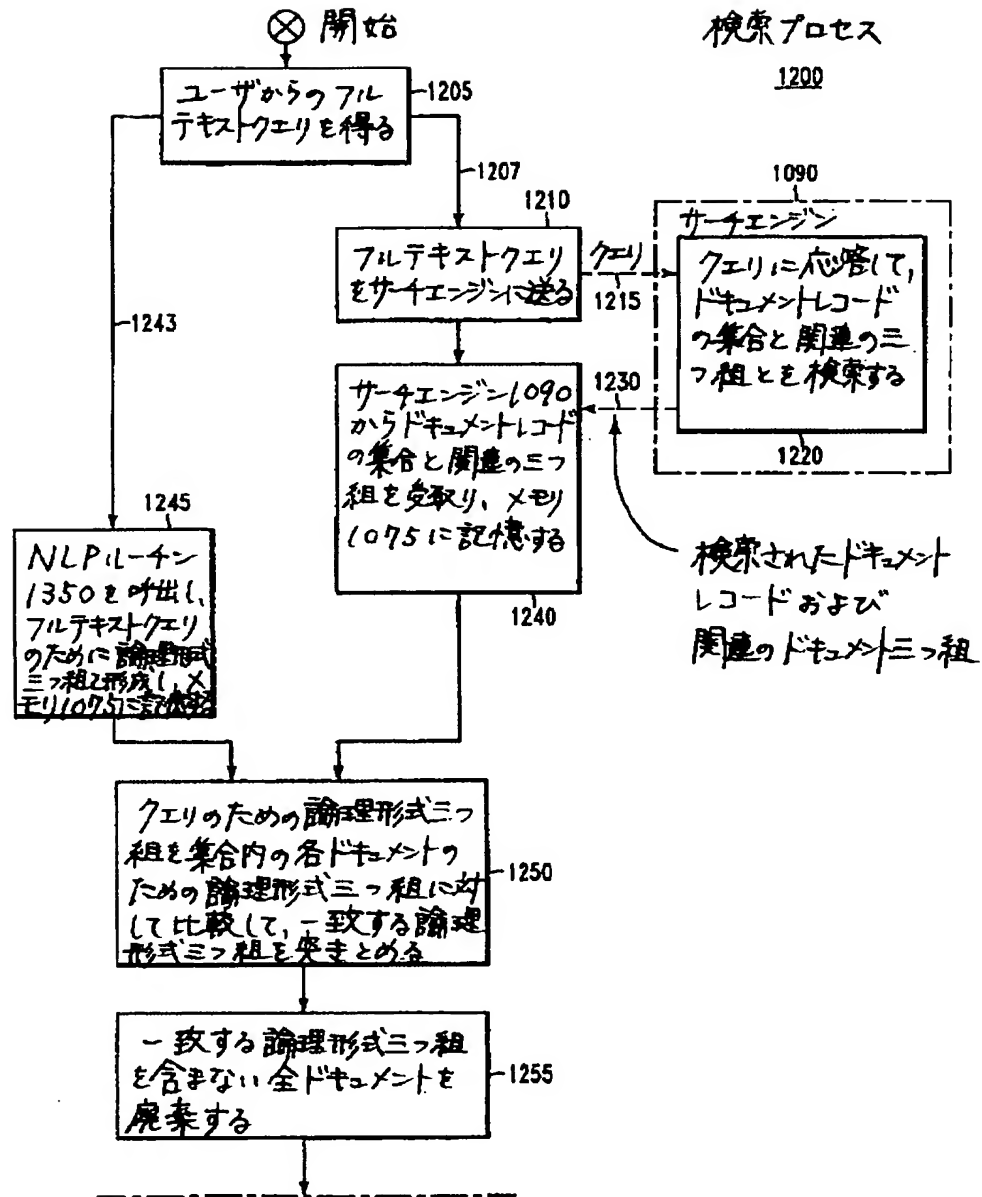


FIG. 12

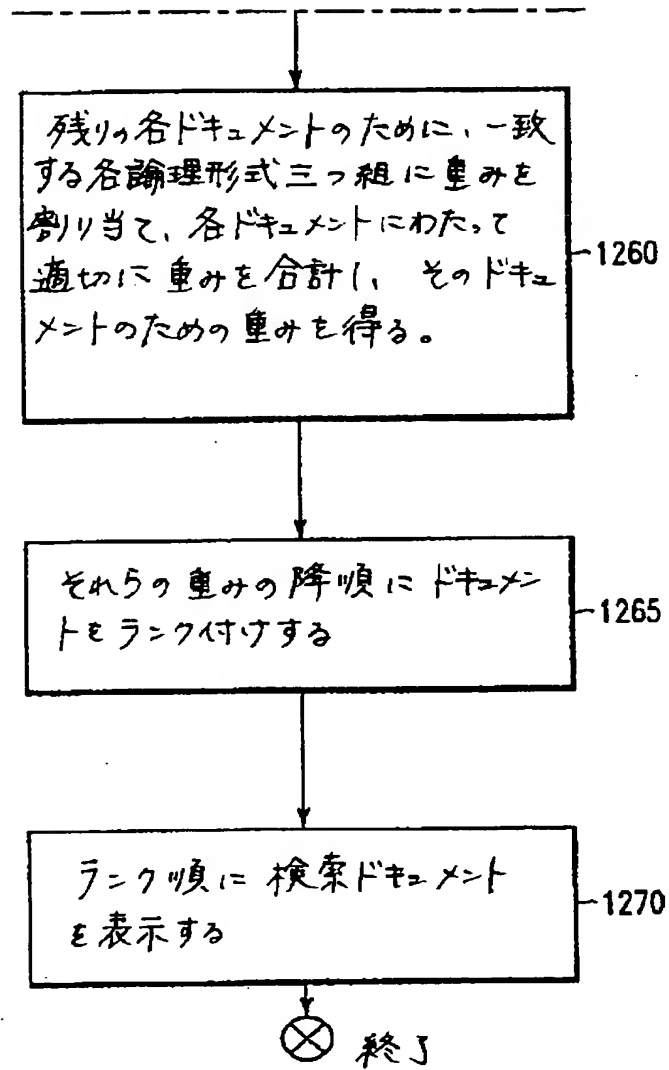
(80)

【図12A】



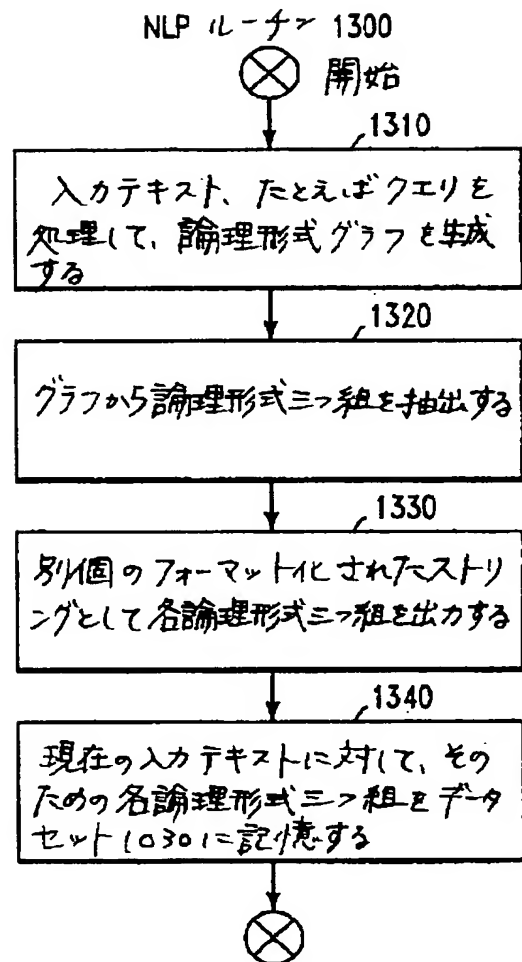
(81)

【図12B】



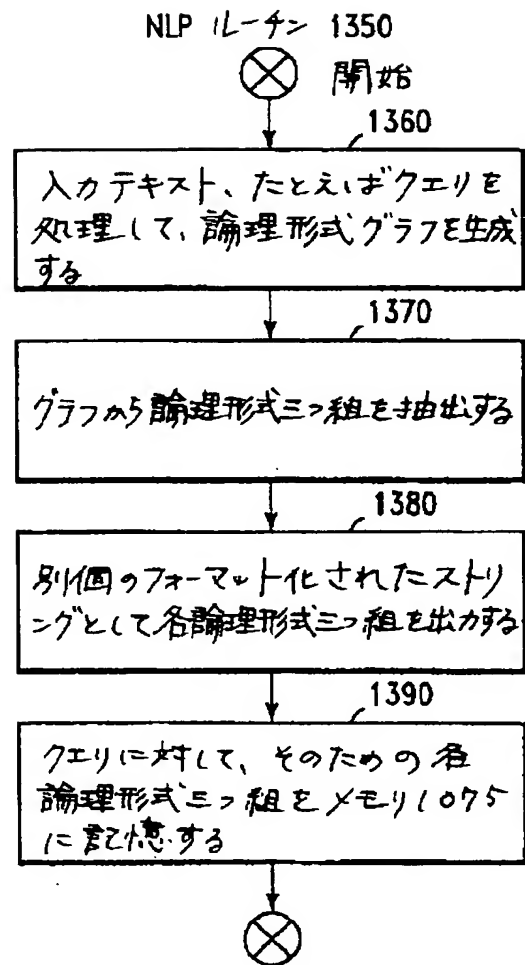
(82)

【図13A】



(83)

【図 13B】



(84)

【手続補正書】

【提出日】平成12年1月25日(2000. 1. 25)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】 記憶されているドキュメントをリポジトリ (10) から検索するための情報検索システム (5) において用いるための装置であって、前記システムは、クエリに回答してそのクエリに関連した複数の記憶されているドキュメントをリポジトリから検索し、出力ドキュメント集合を規定するための検索システム (20) を有し、前記装置は、

プロセッサ (340) と、

実行可能な命令 (600) が記憶されているメモリ (375) とを含み、

プロセッサはメモリに記憶されている命令に回答して、

クエリに回答してそのための第1の論理形式を生じ、第1の論理形式はクエリに関連した語の間の意味的關係を示し、

出力ドキュメント集合内のドキュメントの各別の1つに対して、対応する第2の論理形式を取得し、第2の論理形式は前記1つのドキュメント内の句に関連した語の間の意味的關係を示し、

クエリの第1の論理形式と、出力ドキュメント集合内の複数のドキュメントの各1つのための第2の論理形式との予め定義された関数として、出力ドキュメント集合内の複数のドキュメントをランク付けしてランク順を規定し、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力 (446) として与える、装置。

【請求項2】 各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の1つのドキュメントに関連したレコードである、請求項1に記載の装置。

(85)

【請求項 3】 クエリのための第 1 の論理形式と出力ドキュメント集合内の各別のドキュメントのための第 2 の論理形式との各々はそれぞれ、論理形式グラフ (515, 535, 555, 575)、そのサブグラフ、または論理形式三つ組のリスト (525, 540, 560, 580) である、請求項 2 に記載の装置。

【請求項 4】 プロセッサは記憶されている命令に応答して、
出力ドキュメント集合内のドキュメントの前記各別の 1 つのために、記憶媒体から対応の第 2 の論理形式を読出すか、または
出力ドキュメント集合内の前記各別の 1 つのドキュメントを分析することによって前記対応の第 2 の論理形式を生成する、請求項 3 に記載の装置。

【請求項 5】 前記関数は、前記ドキュメントの 1 つのために、クエリに関連した前記第 1 の論理形式と前記 1 つのドキュメントに関連した前記第 2 の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、プロセッサは記憶されている命令に応答して、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定する、請求項 4 に記載の装置。

【請求項 6】 クエリに関連した前記第 1 の論理形式と出力ドキュメント集合内の任意のドキュメントに関連した前記第 2 の論理形式の任意のものとの間の前記一致は同一の一致である、請求項 5 に記載の装置。

【請求項 7】 ユーザからのクエリ (201) を取得し、出力ドキュメント集合内の複数のドキュメント (203) を前記ランク順に表示するためのクライアントコンピュータ (300) と、

ネットワーク接続 (205, 210, 215) を介してクライアントコンピュータに接続されるサーバ (220) とをさらに含み、前記サーバは前記プロセッサ (340) および前記メモリ (375) を含み、プロセッサはメモリに記憶されている命令 (600) に応答して、

クライアントコンピュータからクエリを取得し、

出力ドキュメントの集合内の前記複数のドキュメントを前記ランク順にクライアントコンピュータに与える、請求項 5 に記載の装置。

(86)

【請求項 8】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1 つ以上の論理形式三つ組の対応の第 1 のリストおよび第 2 のリストを含み、前記第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、2 つの語の間の意味的關係を表わす予め規定された関係とを含む、請求項 5 または 6 に記載の装置。

【請求項 9】 前記 1 つのドキュメントのためのスコアはまた、前記 1 つのドキュメントのための第 2 の論理形式内のノード語、前記 1 つのドキュメント内の前記ノード語の頻度または意味的内容、前記 1 つのドキュメント内の予め規定されたノード語の頻度または意味的内容、前記 1 つのドキュメントのための特定の論理形式三つ組の頻度、もしくは前記 1 つのドキュメントの長さの、予め定められた関数である、請求項 5 または 6 に記載の装置。

【請求項 10】 前記関数は、クエリに関連した論理形式三つ組の少なくとも 1 つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する各論理形式三つ組に割当てられる重みはそれに関連した意味的關係のタイプによって定義される、請求項 8 に記載の装置。

【請求項 11】 プロセッサはメモリに記憶されている命令に応答して、クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定し、

関連した少なくとも 1 つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各 1 つのために、前記一致する論理形式三つ組の各々に関連した意味的關係によって予め規定される重み数値を用いて前記各 1 つのドキュメント内の一致する論理形式三つ組に重み付けして、前記 1 つのドキュメントのための 1 つ以上の重みを形成し、

前記 1 つ以上の重みの関数として前記 1 つのドキュメントのためのスコアを計算し、

(87)

前記ドキュメントの各 1 つをその前記スコアに従ってランク付けしてランク順を規定する、請求項 10 に記載の装置。

【請求項 12】 クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの 1 つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項 8 に記載の装置。

【請求項 13】 クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項 8 に記載の装置。

【請求項 14】 サーチエンジンはクエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各 1 つのために、リポジトリ (10) から記憶されているレコードを検索し、レコードは出力ドキュメント集合内の前記ドキュメントの各 1 つが見出され得る場所を特定する情報を含み、プロセッサはメモリに記憶されている命令とレコードに含まれている情報とに応答して、前記ドキュメントの各 1 つにそのための関連のサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含める、請求項 5、7 または 8 に記載の装置。

【請求項 15】 記憶されているドキュメントをリポジトリ (10) から検索するための情報検索システム (5) において用いるための方法 (600; 1200) であって、前記システムは、クエリに応答してそのクエリに関連した複数の記憶されているドキュメントをリポジトリから検索し、出力ドキュメント集合を規定するための検索システム (20) を有し、前記方法は、

クエリに応答してそのための第 1 の論理形式を生じる (645; 1245) ステップを含み、第 1 の論理形式はクエリに関連した語の間の意味的關係を示し、

出力ドキュメント集合内のドキュメントの各別の 1 つに対して、対応する第 2 の論理形式を取得する (640; 1240) ステップを含み、第 2 の論理形式は前記 1 つのドキュメント内の句に関連した語の間の意味的關係を示し、

クエリの第 1 の論理形式と、出力ドキュメント集合内の複数のドキュメントの各 1 つのための第 2 の論理形式との予め定義された関数 (650, 655, 660; 1250, 1255, 1260) として、出力ドキュメント集合内の複数の

(88)

ドキュメントをランク付けして (665; 1265) ランク順を規定するステップと、

出力ドキュメント集合に関連した複数の記憶されているエントリを前記ランク順に出力として与える (670; 1270) ステップとを含む、方法。

【請求項 16】 各エントリは出力ドキュメント集合内のドキュメントの対応の1つであるか、または前記対応の1つのドキュメントに関連したレコードである、請求項 15 に記載の方法。

【請求項 17】 クエリのための第1の論理形式と出力ドキュメント集合内の各別のドキュメントのための第2の論理形式との各々はそれぞれ、論理形式グラフ (515, 535, 555, 575)、そのサブグラフ、または論理形式三つ組のリスト (525, 540, 560, 580) である、請求項 16 に記載の方法。

【請求項 18】 前記取得するステップは、

出力ドキュメント集合内のドキュメントの前記各別の1つのために、記憶媒体から対応の第2の論理形式を読み出す (1240) か、または

出力ドキュメント集合内の前記各別の1つのドキュメントを分析することによって、前記対応の第2の論理形式を生成する (640) ステップを含む、請求項 17 に記載の方法。

【請求項 19】 前記関数は、前記ドキュメントの1つのために、クエリに関連した前記第1の論理形式と前記1つのドキュメントに関連した前記第2の論理形式の各々との間の予め定められた関係に基づいてスコアを生成し、前記ランク付けするステップは、出力ドキュメント集合内の各ドキュメントに関連したスコアに従って、記憶されているエントリをランク付けしてランク順を規定するステップを含む、請求項 18 に記載の方法。

【請求項 20】 クエリに関連した前記第1の論理形式の任意のものと出力ドキュメント集合内の任意のドキュメントに関連した前記第2の論理形式の任意のものとの間の前記一致は同一の一致である、請求項 19 に記載の方法。

【請求項 21】 システムはクライアントコンピュータをさらに含み、前記方法はクライアントコンピュータにおいて、

(89)

ユーザからのクエリを取得する (605; 1205) ステップと、
出力ドキュメント集合内の複数のドキュメントを前記ランク順に表示する (6
70; 1270) ステップとを含み、
システムはネットワーク接続 (205, 210, 215) を介してクライエ
ントコンピュータに接続されるサーバ (220) をさらに含み、前記方法はサーバ
において、

クライアントコンピュータからクエリを取得するステップと、
出力ドキュメント集合内の前記複数のドキュメントを前記ランク順にクライエ
ントコンピュータに与えるステップとを含む、請求項 19 に記載の方法。

【請求項 22】 前記第 1 の論理形式および前記第 2 の論理形式の各々は 1
つ以上の論理形式三つ組の対応の第 1 のリストおよび第 2 のリストを含み、前記
第 1 のリスト内の前記論理形式三つ組と前記第 2 のリスト内の前記論理形式三つ
組とは各々、それぞれクエリ内のまたは前記ドキュメントの各 1 つの句内の、対
応の論理形式グラフにおいて意味的に関係した 2 つの語の各々の語幹の形と、2
つの語の間の意味的關係を表わす予め規定された関係とを含む、請求項 19 また
は 20 に記載の方法。

【請求項 23】 前記 1 つのドキュメントのためのスコアはまた、前記 1 つ
のドキュメントのための第 2 の論理形式内のノード語、前記 1 つのドキュメント
内の前記ノード語の頻度または意味的内容、前記 1 つのドキュメント内の予め規
定されたノード語の頻度または意味的内容、前記 1 つのドキュメントのための特
定の論理形式三つ組の頻度、もしくは前記 1 つのドキュメントの長さの、予め定
められた関数である、請求項 19 または 20 に記載の方法。

【請求項 24】 前記関数は、クエリに関連した論理形式三つ組の少なくと
も 1 つと同一に一致する、出力ドキュメント集合内の前記複数のドキュメントの
各々に関連した論理形式三つ組にわたってとられた重みの合計であり、一致する
各論理形式三つ組に割当てられる重みはそれに関連した意味的關係のタイプによ
って定義される、請求項 22 に記載の方法。

【請求項 25】 前記ランク付けするステップは、
クエリに関連した論理形式三つ組の任意のものが出力ドキュメント集合内の任

(90)

意のドキュメントに関連した論理形式三つ組の任意のものと一致するか否かを判断して、前記任意のドキュメントに関連した一致する三つ組を規定するステップと、

関連した少なくとも1つの一致する論理形式三つ組を有する前記出力ドキュメント集合内のドキュメントの各1つのために、前記一致する論理形式三つ組の各々に関連した意味的關係によって予め規定される重み数値を用いて前記各1つのドキュメント内の一致する論理形式三つ組に重み付けして、前記1つのドキュメントのための1つ以上の重みを形成するステップと、

前記1つ以上の重みの関数として前記1つのドキュメントのためのスコアを計算するステップと、

前記ドキュメントの各1つをその前記スコアに従ってランク付けしてランク順を規定するステップとを含む、請求項24に記載の方法。

【請求項26】 クエリに関連した前記論理形式三つ組か、または出力ドキュメント集合内の前記ドキュメントの1つに関連した前記論理形式三つ組は、前記語のいずれかの上位語または類義語を含む論理形式三つ組をさらに含む、請求項22に記載の方法。

【請求項27】 クエリに関連した論理形式三つ組の前記任意のものと出力ドキュメント集合内の任意のドキュメントに関連した論理形式三つ組の前記任意のものとの間の前記一致は同一の一致である、請求項22に記載の方法。

【請求項28】 サーチエンジンにおいて、クエリに応答して、出力ドキュメント集合内の前記複数のドキュメントの各1つのために、リポジトリ(10)から記憶されているレコードを検索するステップをさらに含む、レコードは出力ドキュメント集合内の前記ドキュメントの各1つが見出され得る場所を特定する情報を含み、サーバにおいて、レコードに含まれている情報に応答して、前記ドキュメントの各1つにそのための関連するサーバからアクセスし、それをダウンロードし、出力ドキュメント集合内に含めるステップをさらに含む、請求項19、21または22に記載の方法。

【請求項29】 コンピュータで実行可能な命令を記憶し、請求項15に記載のステップを実行するためのコンピュータ読出可能媒体。

(91)

【手続補正 2】

【補正対象書類名】要約書

【補正対象項目名】全文

【補正方法】変更

【補正内容】

【要約】

全体の精度を高めるために、たとえば従来の統計に基づくサーチエンジンのような情報検索エンジン (20) によって検索された結果を処理するために自然言語処理を利用する情報検索システム (5) のための装置およびそれに付随する方法を提供する。具体的には、このようなサーチは最終的に検索されたドキュメントの集合を生む。このような各ドキュメントは次に自然言語処理を受けて論理形式の集合を生じる。このような各論理形式は句内の語間の意味的關係、特に主題と修飾語句との構造を語-関係子-語の態様で符号化する。ユーザが与えるクエリも同様に分析されてそのための対応の論理形式の集合を生み出す。ドキュメントはドキュメントおよびクエリからの論理形式の予め規定された関数としてランク付けされる。具体的には、クエリのための論理形式の集合は、検索されたドキュメントの各々のための論理形式の集合と比較されて両方の集合内のこのような任意の論理形式間の一致を確認する。少なくとも 1 つの一致する論理形式を有する各ドキュメントがヒューリスティックにスコア付けされ、一致する論理形式のための異なる各関係が異なる対応の予め定められた重みを割当てられる。このような各ドキュメントのスコアはたとえば、その独自に一致する論理形式の重みの予め規定された関数である。最後に、保持されたドキュメントがスコアの高い順にランク付けされてその順でユーザに提示される。

(92)

【国際調査報告】

INTERNATIONAL SEARCH REPORT

 Internat'l Application No
 PCT/US 98/09711

 A. CLASSIFICATION OF SUBJECT MATTER
 IPC 6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

 Minimum documentation searched (classification system followed by classification symbols)
 IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category * | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|------------|---|-----------------------|
| A | WO 92 04681 A (GTE LABORATORIES INC) 19 March 1992 see abstract see page 1, line 1 - page 4, line 8 | 1,63 |
| A | EP 0 386 825 A (BSO BURO VOOR SYSTEEMONTWIKKEL) 12 September 1990 see abstract | 1,63 |
| A | WO 96 23265 A (BRITISH TELECOMM ; DAVIES NICHOLAS JOHN (GB); WEEKS RICHARD (GB)) 1 August 1996 see abstract | 1,63 |
| A | WO 95 29452 A (APPLE COMPUTER ; ROSE DANIEL E (US); BORNSTEIN JEREMY J (US); TIENE) 2 November 1995 see abstract | 1,63 |

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

Z* document member of the same patent family

Date of the actual completion of the international search

9 September 1998

Date of mailing of the international search report

16/09/1998

Name and mailing address of the ISA

 European Patent Office, P.B. 5318 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
 Fax: (+31-70) 340-3016

Authorized officer

Katerbau, R

1

(93)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No.

PCT/US 98/09711

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---------------------|----------------------------|---------------------|
| WO 9204681 A | 19-03-1992 | US 5321833 A | 14-06-1994 |
| | | CA 2071485 A | 01-03-1992 |
| | | EP 0497960 A | 12-08-1992 |
| | | JP 5502533 T | 28-04-1993 |
| EP 0386825 A | 12-09-1990 | NL 8900587 A | 01-10-1990 |
| | | CA 2011411 A | 10-09-1990 |
| | | JP 3087975 A | 12-04-1991 |
| | | US 5128865 A | 07-07-1992 |
| WO 9623265 A | 01-08-1996 | AU 4454996 A | 14-08-1996 |
| | | BR 9606931 A | 11-11-1997 |
| | | CA 2210581 A | 01-08-1996 |
| | | CN 1169195 A | 31-12-1997 |
| | | EP 0807291 A | 19-11-1997 |
| | | FI 973080 A | 22-07-1997 |
| | | NO 973372 A | 22-09-1997 |
| WO 9529452 A | 02-11-1995 | US 5724567 A | 03-03-1998 |
| | | AU 2363895 A | 16-11-1995 |

フロントページの続き

- (72) 発明者 コーストン, シモン・エイチ
アメリカ合衆国、98102 ワシントン州、
シアトル、ボイルストン・アベニュー・イ
ー、605、ナンバー・109
- (72) 発明者 ドラン, ウィリアム・ビィ
アメリカ合衆国、98052 ワシントン州、
レッドモンド、ワンハンドレッドアンドフ
ィフティサード・コート・エヌ・イー、
7412
- (72) 発明者 バンダーウェンデ, ルーシー・エイチ
アメリカ合衆国、98008 ワシントン州、
ベレビュ、エヌ・イー・サーティス・スト
リート、16415

F ターム(参考) 5B075 KK07 PP24

【要約の続き】

応の予め定められた重みを割当てられる。このような各
ドキュメントのスコアはたとえば、その独自に一致する
論理形式の重みの予め規定された関数である。最後に、
保持されたドキュメントがスコアの高い順にランク付け
されてその順でユーザに提示される。